

The Phase Transition in Five Point Energy Minimization

Richard Evan Schwartz

November 22, 2016

Contents

I	Introduction and Outline	9
1	Introduction	11
1.1	The Energy Minimization Problem	11
1.2	Main Results	13
1.3	Ideas in the Proof	14
1.4	Computer Code	16
2	Outline of the Proof	19
2.1	Under-Approximation	19
2.2	Divide and Conquer	20
2.3	Symmetrization	23
2.4	Endgame	25
2.5	Computing the Constant	27
II	Under-Approximation	29
3	Outline of the Forcing Lemma	31
3.1	A Four Chapter Plan	31
3.2	The Regular Tetrahedron	31
3.3	The General Approach	32
3.4	Case 1	33
3.5	Case 2	35
3.6	Case 3	36
3.7	Case 4	37
3.8	Have Fun Playing Around	37

4	Polynomial Approximations	39
4.1	Rational Intervals	39
4.2	Interval Polynomials	40
4.3	A Bound on the Power Functions	41
4.4	Approximation of Power Combos	41
5	Positive Dominance	45
5.1	The Weak Positive Dominance Criterion	45
5.2	The Positive Dominance Algorithm	46
5.3	Discussion	50
6	Proof of The Forcing Lemma	51
6.1	Positivity of the Coefficients	51
6.2	Under Approximation: Case 1	53
6.3	Under Approximation: Case 2	55
6.4	Under-Approximation: Case 3	57
6.5	Under Approximation: Case 4	58
III	Divide and Conquer	59
7	The Configuration Space	61
7.1	Normalized Configurations	61
7.2	Compactness Results	62
7.3	The TBP Configurations	64
7.4	Dyadic Blocks	65
7.5	A Technical Result about Dyadic Boxes	68
7.6	Very Near the TBP	68
8	Spherical Geometry Estimates	71
8.1	Overview	71
8.2	Some Results about Circles	72
8.3	The Hull Approximation Lemma	74
8.4	Dot Product Estimates	76
8.5	Disordered Blocks	77
8.6	Irrelevant Blocks	78

9	The Energy Theorem	81
9.1	Main Result	81
9.2	Generalizations and Consequences	82
9.3	A Polynomial Inequality	83
9.4	The Local Energy Lemma	85
9.5	From Local to Global	87
10	The Main Algorithm	91
10.1	Grading a Block	91
10.2	The Divide and Conquer Algorithm	91
10.3	Results of the Calculations	92
10.4	Subdivision Recommendation	93
10.5	Interval Arithmetic	94
10.6	Integer Calculations	96
10.7	A Gilded Approach	97
10.8	A Potential Speed Up	97
10.9	Debugging	98
11	Local Analysis of the Hessian	99
11.1	Outline	99
11.2	Lower Bounds on the Eigenvalues	100
11.3	Taylor's Theorem with Remainder	101
11.4	Variation of the Eigenvalues	102
11.5	The Biggest Term	104
11.6	The Remaining Terms	106
11.7	The End of the Proof	107
IV	Symmetrization	109
12	Preliminaries	111
12.1	Monotonicity Lemma	111
12.2	A Three Step Process	112
12.3	Base and Bows	112
12.4	Proof Strategy	114

13 The Domains	117
13.1 Basic Description	117
13.2 Affine Cubical Coverings	119
13.3 The Coverings for Lemma 1	120
13.4 The Coverings for Lemma 2	123
14 The First Symmetrization	125
14.1 Inequality 1	125
14.2 Inequality 2	127
14.3 Results about Triangles	128
14.4 Inequality 3	132
15 The Second Symmetrization	135
15.1 Inequality 1	135
15.2 Inequality 2	139
15.3 Improved Monotonicity	140
15.4 Inequality 3	141
15.5 Discussion	143
V Endgame	145
16 The Baby Energy Theorem	147
16.1 The Main Result	147
16.2 A General Estimate	149
16.3 A Formula for the Energy	150
16.4 The First Estimate	151
16.5 A Table of Values	152
16.6 The Second Estimate	153
17 Divide and Conquer Again	155
17.1 The Goal	155
17.2 The Ideal Calculation	156
17.3 Floating Point Calculations	157
17.4 Confronting the Power Function	159
17.5 The Rational Calculation	161
17.6 Discussion	162

18 The Tiny Piece	163
18.1 Overview	163
18.2 The Proof Modulo Derivative Bounds	164
18.3 The First Bound	166
18.4 The Second Bound	169
19 The Final Battle	173
19.1 Discussion	173
19.2 Proof Overview	174
19.3 The Critical Interval	175
19.4 The Small Interval	177
19.5 The Long Interval	178
19.6 The End	180

Preface

This monograph is a rigorous computer assisted analysis of the extent to which the triangular bi-pyramid (TBP) is the minimizer, amongst 5-point configurations on the sphere, with respect to a power law potential. Our main result settles a long-standing open question about this, but perhaps some people will object to both the length of the solution and to the computer-assisted nature. On the positive side, the proof divides neatly into 4 modular pieces, and there is a clear idea behind the attack at each step.

Define the Riesz s -potential $R_s(r) = \text{sign}(s) r^{-s}$. We prove there exists a computable number $\varpi = 15.0480773927797\dots$ such that the TBP is the unique minimizer with respect to R_s if and only if $s \in (-2, 0) \cup (0, \varpi)$. The Hebrew letter ϖ is pronounced “shin”. Our result does not say what happens for power law potentials much beyond ϖ , though we do prove that in the tiny interval $(\varpi, 15 + 25/512]$ the minimizer is a pyramid with square base. Thus, we prove the existence of the phase transition that was conjectured to exist in the 1977 paper [MKS] by T. W. Melnyk, O. Knop, and W. R. Smith.

This is still a draft of the monograph. The mathematics is all written down and the computer programs have all been run, but I am not sure if the programs are in their final form. While the programs are now pretty well documented and organized, I might still like to do more in this direction. Meanwhile, I am happy to answer any questions about the code.

I thank Henry Cohn, John Hughes, Abhinav Kumar, Curtis McMullen, Ed Saff, and Sergei Tabachnikov, for discussions related to this monograph. I would specially like to thank Alexander Tumanov for his great idea about this problem, and Jill Pipher for her enthusiasm and encouragement. I thank my wife Brienne Brown for suggesting the name of the phase-transition constant. I thank my family for their forbearance while I worked on this project like Ahab going after the whale.

I thank I.C.E.R.M. for facilitating this project. My interest in this problem was rekindled during discussions about point configuration problems around the time of the I.C.E.R.M. Science Advisory Board meeting in Nov. 2015. I thank the National Science Foundation for their continued support, currently in the form of the Research Grant DMS-1204471. Finally, I thank the Simons Foundation for their support, in the form of a Simons Sabbatical Fellowship.

Part I

Introduction and Outline

Chapter 1

Introduction

1.1 The Energy Minimization Problem

Let S^2 denote the unit sphere in \mathbf{R}^3 and let $P = \{p_1, \dots, p_n\}$ be a finite list of distinct points on S^2 . Given some function $f : (0, 2] \rightarrow \mathbf{R}$ we can compute the total f -potential

$$\mathcal{E}_f(P) = \sum_{i < j} f(\|p_i - p_j\|). \quad (1.1)$$

For fixed f and n , one can ask which configuration(s) minimize $\mathcal{E}_f(P)$.

For this problem, the energy functional $f = R_s$, where

$$R_s(r) = \text{sign}(s)r^{-s}, \quad (1.2)$$

is a natural one to consider. When $s > 0$, this is called the *Riesz potential*. When $s < 0$ this is called the *Fejes-Toth potential*. The case $s = 1$ is specially called the *Coulomb potential* or the *electrostatic potential*. This case of the energy minimization problem is known as *Thomson's problem*. See [Th]. The case of $s = -1$, in which one tries to maximize the sum of the distances, is known as *Polya's problem*.

There is a large literature on the energy minimization problem. See [Fö] and [C] for some early local results. See [MKS] for a definitive numerical study on the minimizers of the Riesz potential for n relatively small. The website [CCD] has a compilation of experimental results which stretches all the way up to about $n = 1000$. The paper [SK] gives a nice survey of results, with an emphasis on the case when n is large. See also [RSZ]. The paper

[BBCGKS] gives a survey of results, both theoretical and experimental, about highly symmetric configurations in higher dimensions.

When $n = 2, 3$ the problem is fairly trivial. In [KY] it is shown that when $n = 4, 6, 12$, the most symmetric configurations – i.e. vertices of the relevant Platonic solids – are the unique minimizers for all R_s with $s \in (-2, \infty) - \{0\}$. See [A] for just the case $n = 12$ and see [Y] for an earlier, partial result in the case $n = 4, 6$. The result in [KY] is contained in the much more general and powerful result [CK, Theorem 1.2] concerning the so-called sharp configurations.

The case $n = 5$ has been notoriously intractable. There is a general feeling that for a wide range of energy choices, and in particular for the Riesz potentials, the global minimizer is either the triangular bi-pyramid or else some pyramid with square base. The *triangular bi-pyramid* (TBP) is the configuration of 5 points with one point at the north pole, one at the south pole, and three arranged in an equilateral triangle around the equator. The pyramids with square base come in a 1-parameter family. Following Tumanov [T] we call them *four pyramids* (FPs).

[HS] has a rigorous computer-assisted proof that the TBP is the unique minimizer for R_{-1} (Polya's problem) and my paper [S1] has a rigorous computer-assisted proof that the TBP is the unique minimizer for R_1 (Thomson's problem) and R_2 . The paper [DLT] gives a traditional proof that the TBP is the unique minimizer for the logarithmic potential.

In [BHS, Theorem 7] it is shown that, as $s \rightarrow \infty$, any sequence of 5-point minimizers w.r.t. R_s must converge (up to rotations) to the FP having one point at the north pole and the other 4 points on the equator. In particular, the TBP is not a minimizer w.r.t. R_s when s is sufficiently large. In 1977, T. W. Melnyk, O. Knop, and W. R. Smith, [MKS] conjectured the existence of the phase transition constant, around $s = 15.04808$, at which point the TBP ceases to be the minimizer w.r.t. R_s .

Define

$$G_k(r) = (4 - r^2)^k, \quad k = 1, 2, 3, \dots \quad (1.3)$$

In [T], A. Tumanov gives a traditional proof of the following result.

Theorem 1.1.1 (Tumanov) *Let $f = a_1 G_1 + a_2 G_2$ with $a_1, a_2 > 0$. The TBP is the unique global minimizer with respect to f . Moreover, a critical point of f must be the TBP or an FP.*

As an immediate corollary, the TBP is a minimizer for G_1 and G_2 . Tumanov points out that these potentials do not have an obvious geometric

interpretation, but they are amenable to a traditional analysis. Tumanov also mentions that his result might be a step towards proving that the TBP minimizes a range of power law potentials. He makes the following observation.

Tumanov's Observation: If the TBP is the unique minimizer for G_2 , G_3 and G_5 , then the TBP is the unique minimizer for R_s provided that $s \in (-2, 2] - \{0\}$.

Tumanov does not offer a proof for his observation, but we will prove related results in this monograph. Tumanov also remarks that it seems hard to establish that the TBP is the minimizer w.r.t G_3 and G_5 . The family of potentials $\{G_k\}$ behaves somewhat like the Riesz potentials. For instance, the TBP is not the global minimizer w.r.t. G_7 . I checked that the TBP is not the global minimizer for G_8, \dots, G_{100} , and I am sure that this fact remains true for $k > 100$ as well.

1.2 Main Results

The idea of this monograph is to combine the divide-and-conquer approach in [S1] with (elaborations of) Tumanov's observation (and other tricks) to rigorously establish the phase transition explicitly conjectured in [MKS]. We will show that the TBP is the unique global minimizer for

$$G_3, \quad G_4, \quad G_5, \quad G_6, \quad G_5 - 25G_1, \quad G_{10} + 28G_5 + 102G_2,$$

and that most configurations (in a sense made precise in the next chapter) have higher $G_{10} + 13G_5 + 68G_2$ energy than the TBP. We will then use these facts to establish the following two results.

Theorem 1.2.1 *Let $s \in (-2, 0)$ be arbitrary. Amongst all 5-point configurations on the sphere, the TBP is the unique global minimizer w.r.t. R_s .*

Theorem 1.2.2 *Let $s \in (0, 15 + 25/512]$ be arbitrary. Amongst all 5-point configurations on the sphere, either the TBP or some FP is the unique global minimizer w.r.t. to R_s .*

We chose $15 + 25/512 = 15.04809\dots$ because this is a dyadic rational that is very slightly larger than the cutoff in [MKS]. The above two results combine with some elementary analysis to establish our main result.

Theorem 1.2.3 (Main) *There exists a computable number*

$$\varpi = 15.0480773927797\dots$$

such that the TBP is the unique global minimizer w.r.t. the R_s energy if and only if $s \in (-2, 0) \cup (0, \varpi)$.

Our results say nothing about what happens beyond $s = 15 + 25/512$ (though our Symmetrization Lemma in §2.3 perhaps suggests a path onward; see the discussion in §19.1.) Even so, the Main Theorem is the definitive result about the extent that the TBP minimizes power law potentials. In particular, it establishes the existence of the phase transition constant which has been conjectured to exist for many decades.

The constant ϖ is probably some transcendental number that is unrelated to all other known constants of nature and mathematics. So, saying that an ideal computer can compute ϖ to arbitrary precision is the best we can do. In §2.5 we will sketch an algorithm which computes ϖ to arbitrary precision.

1.3 Ideas in the Proof

In the next chapter we will give a detailed outline of the proofs of the results above. Here we make some general comments about the strategy.

We are interested in searching through the moduli space of all 5-point configurations and eliminating those which have higher energy than the TBP. We win if we eliminate everything but the TBP. Assuming that the TBP really is the minimizer for a given energy function f , it is probably the case that most configurations are not even close to the TBP in terms of f -energy. So, in principle, one can eliminate most of the configuration space just by crude calculations. If this elimination procedure works well, then what is left is just a small neighborhood B of the TBP. The TBP is a critical point for \mathcal{E}_f , by symmetry, and with luck the Hessian of \mathcal{E}_f is positive definite throughout B . In this case, we can say that the TBP must be the unique global minimizer.

Most of our proof involves implementing such an elimination scheme. We use a divide-and-conquer approach. We normalize so that $(0, 0, 1)$ is one of the points of the configuration, and then we use stereographic projection to move this point to ∞ and the other 4 points to \mathbf{R}^2 . Stereographic projection

is the map

$$\Sigma(x, y, z) = \left(\frac{x}{1-z}, \frac{y}{1-z} \right). \quad (1.4)$$

We call these points *stereographic coordinates* for the configuration. Stereographic coordinates give the moduli space a natural flat structure, making divide-and-conquer algorithms easy to manage. Our basic object is a *block*, a rectangular solid in the moduli space. The main mathematical feature of the paper is a result which gives a lower bound on the energy of any configuration in a block based on the potentials of the configurations corresponding to the vertices, and an error term. The Energy Theorem (Theorem 9.2.1) is our main result along these lines. The Energy Theorem is related to the same kind of result we had in [S1] but the result here is cleaner because we are dealing with polynomial expressions. We state the Energy Theorem specifically for the energy potentials G_k mentioned above, but in principle it should work much more broadly.

It is hard to over-emphasize the importance of having an efficient error term in the Energy Theorem. This makes the difference between a feasible calculation and one which would outlast the universe. I worked quite hard on making the Energy Theorem simple, efficient, and pretty sharp. Another nice feature of the Energy Theorem is that every quantity involved in it is a rational function of the vertices of the block. Thus, at least in principle, we could run all our computer programs using exact integer arithmetic. Such integer calculations seemed too slow, however. So, I implemented the calculations using interval arithmetic. Since the terms in the Energy Theorem are all rational, our calculations only involve the operations plus, minus, times, divide. Such operations are controlled by the IEEE standards, so that interval arithmetic is easy to implement. In contrast, if we had to rigorously bound expressions involving the power function¹ we would be in deeper and murkier waters computationally.

So far we have discussed one energy function at a time, but we are interested in a 1-parameter family of power laws and we can only run our program finitely many times. This is where Tumanov's observation comes in. For instance, if we just run our program on G_3 and G_5 (which we do) we pick up all the R_s functions for $s \in (-2, 0) \cup (0, 2]$. After a lot of experimenting I found variants of Tumanov's result which cover the rest of the exponent

¹Towards the end of the monograph, in §16, we are forced to confront the power function directly for a very limited calculation, but we will reduce everything to the basic arithmetic operations in this case.

range $[0, 15 + 25/512]$. Our main result along these lines, which is really a compilation of results, is the Forcing Lemma. See §2.1.

Since $\varpi < 15 + 25/512$ something must go wrong with our approach above. What happens is the TBP is not quite the minimizer with respect to one of the magic energy functions,

$$G_{10}^{\sharp} = G_{10} + 13G_5 + 68G_2. \quad (1.5)$$

However we can still eliminate most of the configuration space. What remains is a small region **SMALL** of mystery configurations. See 2.2.

We deal with the configurations in **SMALL** using a symmetrization construction that, so to speak, is neutral with respect to the TBP/FP dichotomy. The symmetrization construction, which only decreases energy in a small but critical range, somehow makes the TBP and the FPs work together rather than as competitors. We explain the construction in §2.3 and elaborate on the philosophical reason why it works.

1.4 Computer Code

The computer code involved in this paper is publicly available by download from my website,

<http://www.math.brown.edu/~res/Java/TBP.tar>

All the algorithms are implemented in Java. We also have a small amount of Mathematica code needed for the analysis of the Hessians done in §11, and also for the handful of evaluations done in §19. In both cases, we use Mathematica mostly for the sake of convenience, and the calculations are so straightforward and precisely described that I didn't see the need to implement them in Java. The reader who knows Mathematica or (say) Matlab or Sage could re-implement these calculations in about an hour.

The (non-Mathematica) programs come with graphical user interfaces, and each one has built in documentation and debuggers. The interfaces let the user watch the programs in action, and check in many ways that they are operating correctly. The computer programs are distributed in 4 directories, corresponding to the 4 main phases of the proof.

The overall computer program is large and complicated, but mostly because of the elaborate graphical user interfaces. These interfaces are ex-

tremely useful for debugging purposes, and also they give a good feel for the calculations. However, they make the code harder to inspect. To remedy this situation, I have made separate directories which just contain the code which goes into the computational tests. In these directories all the superfluous code associated to the interfaces has been stripped away. The reader can launch the tests either from the interface or from a command line in these stripped-down directories. The tests do the same thing in each case, though with the graphical interface there is more room for varying the conditions of the test (away from what is formally needed) for the purposes of experimentation and control.

I think that a competent programmer would be able to recreate all the code that goes into the tests, either by reading the descriptions in this monograph or else by inspecting the code itself.

Chapter 2

Outline of the Proof

In this chapter we outline the proof of the main results. The proof has 4 main ingredients, Under-Approximation, Divide and Conquer, Symmetrization, and Endgame. We discuss these ingredients in turn. To avoid irritating trivialities, we consider two configurations the same if they are isometric. Given the function f and a 5-point configuration X we often write $f(X)$ in place of $\mathcal{E}_f(X)$ for ease of notation.

2.1 Under-Approximation

Let $I \subset \mathbf{R}$ denote an interval, which we think of an interval of power law exponents. We say that a triple $(\Gamma_2, \Gamma_3, \Gamma_4)$ of potentials is *forcing* on the interval I if the following implication holds for any configuration 5-point configuration X :

$$\Gamma_i(X) > \Gamma_i(TBP) \quad i = 2, 3, 4 \quad \implies \quad R_s(X) > R_s(TBP) \quad \forall s \in I. \quad (2.1)$$

Recall that $G_k(r) = (4 - r^2)^k$. Tumanov's observation is that (G_2, G_3, G_5) is forcing on $(-2, 0)$ and $(0, 2]$. Tumanov does not supply a proof for his observation, but we will prove a more extensive related result. Consider the following functions:

$$\begin{aligned} G_5^b &= G_5 - 25G_1 \\ G_{10}^\sharp &= G_{10} + 13G_5 + 68G_2 \\ G_{10}^{\sharp\sharp} &= G_{10} + 28G_5 + 102G_2 \end{aligned} \quad (2.2)$$

In §3-6 we prove the following result.

Lemma 2.1.1 (Forcing) *The following is true.*

1. (G_2, G_4, G_6) is forcing on $(0, 6]$.
2. $(G_2, G_5, G_{10}^{\#\#})$ is forcing on $[6, 13]$.
3. $(G_2, G_5^\flat, G_{10}^\#)$ is forcing on $[13, 15.05]$.
4. (G_2, G_3, G_5) is forcing on $(-2, 0)$.

Remark:

- (i) We list the negative case last because it is something of a distraction. The triple (G_2, G_3, G_5) is also forcing on $(0, 2]$, as Tumanov observed, but we ignore this case because it is redundant.
- (ii) These kinds of under-approximation results, in some form, are used in many papers on energy minimization. See [BDHSS] for very general ideas like this.

2.2 Divide and Conquer

Using the divide-and-conquer algorithm mentioned in the introduction, together with a local analysis of the Hessian, we give a rigorous, computer-assisted proof of the following result.

Theorem 2.2.1 (Big) *The TBP is the unique global minimizer w.r.t.*

$$G_3, G_4, G_5^\flat, G_6, G_{10}^{\#\#}.$$

We note that G_5 is a positive combination of G_1 and G_5^\flat , so G_5 also satisfies the conclusion of the Big Theorem. The Big Theorem fails for G_7, G_8, \dots . This failure is what led me to search for more complicated combinations like $G_{10}^\#$.

Combining the Big Theorem and the Forcing Lemma, we get

Corollary 2.2.2 *The TBP is the unique global minimizer w.r.t. R_s for all $s \in (-2, 0) \cup (0, 13]$.*

Corollary 2.2.2 contains Theorem 1.2.1. For Theorem 1.2.2 we just have to deal with the exponent interval $(13, 15 + 25/512]$. The TBP is not the global minimizer for $G_{10}^\#$, but we can still squeeze information out of this function.

We work in stereographic coordinates, as in Equation 1.4. Precisely, we have the correspondence:

$$V_0, V_1, V_2, V_3, (0, 0, 1) \in S^2 \iff p_0, p_1, p_2, p_3 \in \mathbf{R}^2, \quad p_k = \Sigma(V_k). \quad (2.3)$$

We write $p_k = (p_{k1}, p_{k2})$. We always normalize so that p_0 lies in the positive x -axis and $\|p_0\| \geq \|p_k\|$ for $k = 1, 2, 3$. Let **SMALL** denote those 5-point configurations which are represented by 4-tuples $\{p_0, p_1, p_2, p_3\}$ such that

1. $\|p_0\| \geq \|p_k\|$ for $k = 1, 2, 3$.
2. $512p_0 \in [433, 498] \times [0, 0]$.
3. $512p_1 \in [-16, 16] \times [-464, -349]$.
4. $512p_2 \in [-498, -400] \times [0, 24]$.
5. $512p_3 \in [-16, 16] \times [349, 364]$.

This domain is pretty tight. We tried hard to get as far away from the TBP configuration as possible.

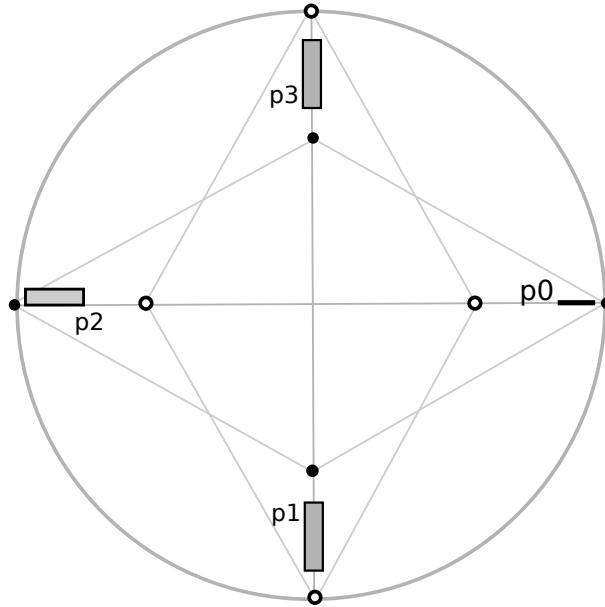


Figure 2.1: The sets defining **SMALL**.

Figure 2.1 shows a picture of the sets corresponding to the definition of **SMALL**. The grey circle is the unit circle. Note that $TBP \notin \mathbf{SMALL}$. The 4 black dots and the 4 white dots are the two nearby normalized TBP configurations. The inner dots are $\sqrt{3}/3$ units from the origin.

Theorem 2.2.3 (Small) *If $G_{10}^\sharp(X) \leq G_{10}^\sharp(TBP)$ then either $X = TBP$ or $X \in \mathbf{SMALL}$.*

Combining this result with the Forcing Lemma, we get

Corollary 2.2.4 *Let X be any 5-point configuration. Let*

$$s \in \left[13, 15 + \frac{25}{512}\right].$$

If $R_s(X) \leq R_s(TBP)$ then either $X = TBP$ or $X \in \mathbf{SMALL}$.

We cannot plot the 7-dimensional region **SMALL**, but we can plot a canonical slice. Let **K4** denote the 2-dimensional set of configurations whose stereographic projections are rhombi with points in the coordinate axes. (Here **K4** stands for “Klein 4 symmetry”.) Let

$$\mathbf{SMALL4} = \mathbf{K4} \cap \mathbf{SMALL}. \quad (2.4)$$

This is the set of configurations (p_0, p_1, p_2, p_3) such that

$$-p_2 = p_0 = (x, 0), \quad -p_1 = p_3 = (0, y), \quad x \geq y, \quad x, y \in \left[\frac{348}{512}, \frac{495}{512}\right].$$

The point $(x, y) = (1, 1/\sqrt{3})$, outside our slice, represents the TBP. The region **SMALL4** is the unshaded region below the gray diagonal in Figure 2.2. Figure 2.2 shows a plot of G_{10}^\sharp as a function of x and y . The darkly shaded region is where $G_{10}^\sharp(x, y) \leq G_{10}^\sharp(TBP)$. Note that this dark region is contained in the union of **SMALL4** and its reflection across the diagonal. This plot doesn’t prove anything but it serves as a sanity check on our calculations.

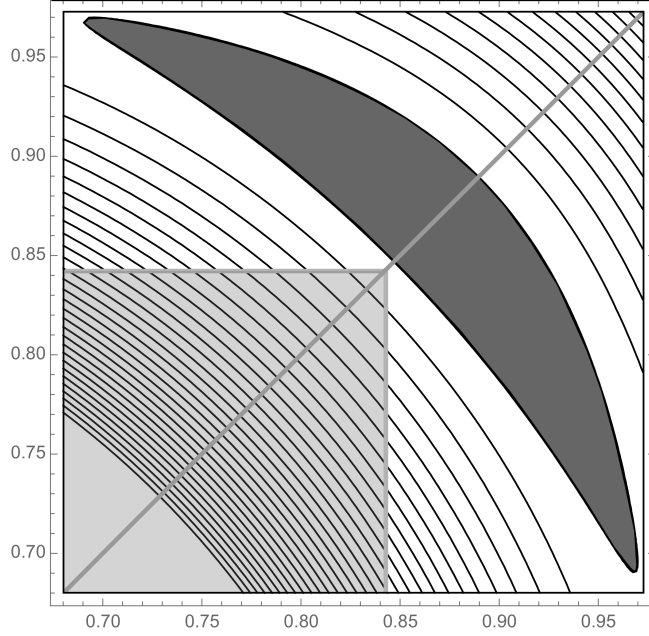


Figure 2.2: A plot inside **SMALL4**.

It remains to deal with configurations in **SMALL**. We take an approach which crucially involves the slice **SMALL4**.

2.3 Symmetrization

Here is the heart of the matter: We will describe an energy decreasing retraction from **SMALL** into **K4**. For some other choices of the number of points – e.g. 6 points – there is a configuration – e.g. the octahedron – whose global symmetry dominates over all other configurations. For 5 points there are two kinds of fairly symmetric configurations, the TBP and FPs, whose influences compete and weaken each other. However, if we align these configurations so that parts of their symmetry groups coincide – i.e., both configurations are members of **K4** – then this aligned symmetry dominates everything nearby. The existence of an energy decreasing symmetrization is a precise formulation of this idea. Philosophically speaking, we make progress through the critical region by aligning the competitors and making them unite their forces.

We consider the following map from **SMALL** to **K4**. We start with the

configuration X having points $(p_1, p_2, p_3, p_4) \in \mathbf{SMALL}$. We let

$$(p'_1, p'_2, p'_3, p'_4)$$

be the configuration which is obtained by rotating X about the origin so that p'_0 and p'_2 lie on the same horizontal line, with p'_0 lying on the right. This is just a very slight rotation. We then define

$$-p_2^* = p_0^* = \left(\frac{p'_{01} - p'_{21}}{2}, 0 \right), \quad -p_3^* = p_1^* = \left(0, \frac{p'_{12} - p'_{32}}{2} \right) \quad (2.5)$$

The points $(p_0^*, p_1^*, p_2^*, p_3^*)$ define the symmetrized configuration X^* . After a brutal struggle, I was finally able to prove the following result:

Lemma 2.3.1 (Symmetrization) *Let $s \in [12, 15 + 25/512]$ and suppose that $X \in \mathbf{SMALL}$. Then $R_s(X^*) \leq R_s(X)$ with equality iff $X = X^*$.*

Combining the Symmetrization Lemma and the Small Theorem, we get the following result:

Lemma 2.3.2 (Dimension Reduction) *Let $s \in [13, 15 + 25/512]$. Suppose that $R_s(X) \leq R_s(TBP)$ and $X \neq TBP$. Then $X \in \mathbf{SMALL4}$.*

The Dimension Reduction Lemma practically finishes the proofs of the Main Theorems. It leaves us with the exploration of a 2-dimensional rectangle in the configuration space. Before we get to the endgame, I want to discuss the Symmetrization Lemma in some detail.

One strange thing is that the map $X \rightarrow X^*$ is not clearly related to spherical geometry. Rather, it is a linear projection with respect to the stereographic coordinates we impose on the moduli space. On the other hand, the beautiful formula in Equation 14.8 suggests (to me) that there is something deeper going on. I found the map $X \rightarrow X^*$ experimentally, after trying many other symmetrization schemes, some related to the geometry of S^2 and some not. You might say that the map defined above is the winner of a Darwinian competition.

Experimentally, the conclusion of the Symmetrization Lemma seems to be more robust than our result suggests. It seems to hold in a domain somewhat larger than **SMALL** and for all $s \geq 2$. However, our proof is quite delicate and uses the precise domain **SMALL** crucially.

Many ideas go into the proof of the Symmetrization Lemma, some geometric, some analytic, and some algebraic. I'd say that the key idea is to use algebra to implement some geometric ideas. For instance, one step in proving a geometric inequality about triangles involves showing that some polynomials in 5 variables, having about 3000 terms each, are positive on unit cube in R^5 . These functions turn out to be positive dominant in the sense of §5, and so we have a short positivity certificate available. Naturally, I would prefer a clean and simple proof of the Symmetrization Lemma, but the crazy argument I give is the best I can do right now.

2.4 Endgame

The Dimension Reduction Lemma makes the endgame fairly routine. What remains is the exploration of the 2-dimensional rectangle **SMALL4**. It is useful to introduce a much smaller set. Define

$$I = (a, b), \quad a = \frac{55}{64}, \quad b = \frac{56}{64}. \quad (2.6)$$

Let **TINY4** be denote the square consisting of configurations such that

$$-p_2, p_0 \in I \times \{0\}, \quad -p_1, p_3 \in \{0\} \times I. \quad (2.7)$$

We capture ϖ in a tiny parameter interval $[\alpha, \beta]$ where

$$\alpha = 15 + \frac{24}{512}, \quad \beta = 15 + \frac{25}{512}. \quad (2.8)$$

We prove the following result by calculations which are similar to those which prove the Big Theorem and the Small Theorem:

Lemma 2.4.1 *Suppose that one of the following is true:*

1. $s \in [13, \alpha]$ and $X \in \mathbf{SMALL4}$.
2. $s \in [\alpha, \beta]$ and $X \in \mathbf{SMALL4} - \mathbf{TINY4}$.

Then $R_s(X) > R_s(TBP)$.

Combining Lemma 2.4.1 with the Dimension Reduction Lemma, we see that the TBP must be the unique global minimizer w.r.t. R_s when $s \in [13, \alpha]$ and that the global minimizer when $s \in [\alpha, \beta]$ is either the TBP or some member of **TINY4**.

To finish the proof, we use another symmetrization trick. The plane is foliated by parabolas having equations of the form

$$\gamma(t) = (t_0 + t - \lambda t^2, t_0 - t - \lambda t^2), \quad \lambda = 2. \quad (2.9)$$

This foliation induces a natural retraction from \mathbf{R}^2 to the main diagonal. We map each point on $\gamma(t)$ to $\gamma(0)$. When we identify **TINY4** with a subset of \mathbf{R}^2 by taking the coordinates (p_{01}, p_{32}) , we get a retraction from **TINY4** onto the subset consisting of pyramids with square base. We call this map the *parabolic retraction*.

Lemma 2.4.2 *Let $s \in [\alpha, \beta]$ and $X \in \mathbf{TINY4}$. Let X^* denote the image of X under the parabolic retraction. Then $R_s(X^*) \leq R_s(X)$, with equality iff $X = X^*$.*

Remark: Lemma 2.4.2 is quite delicate. The analogous result fails if we replace $\lambda = 2$ by $\lambda = 1$ or $\lambda = 4$. I found the parabolic retraction after a lot of trial and error.

Lemma 2.4.2 finishes the proof of Theorem 1.2.2. We actually know more at this point:

1. The TBP is the unique minimizer w.r.t R_s when $s \in (0, \alpha]$.
2. When $s \in [\alpha, \beta]$ any minimizer w.r.t. R_s is either the TBP or an FP.
3. When $s = \beta$ some FP is the unique global minimizer.

The third statement follows from Lemma 2.2.3, Lemma 2.4.2, and a check that a certain FP has lower energy than the TBP at the parameter β . In §19 we deduce the Main Theorem from these two three statements and some straightforward tricks with calculus.

2.5 Computing the Constant

We close this chapter with a sketch of how an ideal computer would compute \mathfrak{w} . The R_s energy of the TBP is

$$R(s) = 6 \times 2^{-s/2} + 3 \times 3^{-s/2} + 4^{-s/2}. \quad (2.10)$$

Holding s fixed, let $\phi(t)$ be the R_s energy of the FP with vertices

$$(\pm\sqrt{1-t^2}, 0, t), \quad (0, \pm\sqrt{1-t^2}, t), \quad (0, 0, 1). \quad (2.11)$$

We have

$$\phi(t) = 2\left(4 - 4t^2\right)^{-s/2} + 4\left(2 - 2t\right)^{-s/2} + 4\left(2 - 2t^2\right)^{-s/2}. \quad (2.12)$$

Let $I = (-3/16, -1/16)$. One check that for each $s \in [15, 15.1]$, that:

- ϕ has it minimum in I .
- $|\phi'(t)| < 1/10$ in I . That is, ϕ is $(1/10)$ -Lipschitz.
- $\phi''(t) > 0$ on I . That is, ϕ is convex on I .

Since ϕ is convex in I , one can compute its minimum value $Q(s)$ to arbitrary precision using a simple iterative algorithm. Namely,

- Let $a_{0,0} = -3/16$ and $a_{0,1} = -1/16$.
- Let $b_k = (a_{k,0} + a_{k,1})/2$.
- If $\phi'(b_k) > 0$ let $a_{k+1,0} = a_{k,0}$ and $a_{k+1,1} = b_k$.
- If $\phi'(b_k) < 0$ let $a_{k+1,0} = b_k$ and $a_{k+1,1} = a_{k,1}$.

If $\phi'(b_k) = 0$ for some k then $Q(s) = \phi(b_k)$. Otherwise $\{b_k\}$ converges and $Q(s) = \lim \phi(b_k)$.

The above iteration is a step in a similar algorithm which computes \mathfrak{w} . First I describe the *comparison step*. Assuming that $s \neq \mathfrak{w}$ then one of two things is true:

1. There is some k for which $\phi(b_k) < R(s)$. In this case $s > \mathfrak{w}$.

2. There is some k for which

$$\phi(b_k) - R(s) > \frac{a_{k,1} - a_{k,0}}{10}.$$

In this case $Q(s) > R(s)$ and $s < \mathfrak{w}$.

Here is the algorithm which computes \mathfrak{w} .

- Set $a_{0,0} = 15 + (1/32)$ and $a_{0,1} = 15 + (3/32)$.
- Let $b_k = (a_{k,0} + a_{k,1})/2$.
- If $b_k < \mathfrak{w}$ let $a_{k+1,0} = b_k$ and $a_{k+1,1} = a_{k,1}$.
- If $b_k > \mathfrak{w}$ let $a_{k+1,0} = a_{k,0}$ and $a_{k+1,1} = b_k$.

Then $a_{k,0} \rightarrow \mathfrak{w}$ from below and $a_{k,1} \rightarrow \mathfrak{w}$ from above. There is one *caveat*: If $\mathfrak{w} = b_k$ for some k , the algorithm will simply run forever at the comparison step and we will have detected an unspeakably improbable coincidence.

A rigorous implementation of this algorithm would be tedious, on account of the non-integer power operations, but a lazy floating-point implementation yields

$$\mathfrak{w} = 15.0480773927797\dots$$

I dropped off the last few digits to be sure about floating point error.

Part II

Under-Approximation

Chapter 3

Outline of the Forcing Lemma

3.1 A Four Chapter Plan

The next 4 chapters are devoted to the proof of the Forcing Lemma. In this chapter we will explain the ideas in the proof and leave certain details unproved. In place of proofs, we will show computer plots. In §4 and §5 we develop the machinery needed to take care of the details. Once the machinery is assembled we will finish the proof of the Forcing Lemma in §6. We split things up this way so that the reader can first get a feel for the Forcing Lemma through numerical evidence and pictures. Perhaps the reader will appreciate the need for the technical machinery we use to convert the numerical evidence into rigorous proofs. (If not, then I would welcome a shorter proof!)

3.2 The Regular Tetrahedron

We first explain a toy version of the Forcing Lemma, which works for the regular tetrahedron. I learned this from the paper [CK], and I think that the idea goes back to Yudin [Y]. It is the source for many results about energy minimization, including the Forcing Lemma.

We set $\Gamma_0 = G_0$, the function that is identically one, and $\Gamma_1 = G_1$. Again, we have

$$G_k(r) = (4 - r^2)^k. \quad (3.1)$$

Lemma 3.2.1 *A configuration of N points on S^2 has minimal Γ_1 energy if and only if the center of mass of the configuration is the origin.*

Proof: For vectors $V, W \in S^2$ we have

$$\Gamma_1(V, W) = 4 - (V - W) \cdot (V - W) = 2 + 2V \cdot W. \quad (3.2)$$

The total energy of a N -point configuration $\{V_i\}$ is

$$\sum_{i < j} 2 + 2V_i \cdot V_j = N(N - 2) + \sum_{i, j} V_i \cdot V_j = N(N - 2) + \left\| \sum_i V_i \right\|^2.$$

Obviously, this is minimized if and only if $\sum V_i = 0$. ♠

Corollary 3.2.2 *Suppose that $\phi : (0, 4] \rightarrow \mathbf{R}$ is a convex decreasing function and $R(r) = \phi(r^2)$. Then the regular tetrahedron is the unique minimizer w.r.t. R amongst all 4-point configurations.*

Proof: The distance between any pair of points in the regular tetrahedron is $\sqrt{8/3}$. Since Γ_1 is linear in r^2 and the graph has negative slope, we can find a combination $\Gamma = a_0\Gamma_0 + a_1\Gamma_1$, with $a_1 > 0$, such that $\Gamma(r) \leq R(r)$ with equality if and only if $r = \sqrt{8/3}$. Supposing that X is some configuration and X_0 is the regular tetrahedron, we have

$$R(X) \geq \Gamma(X) = a_0 + a_1\Gamma_1(X) \geq a_0 + a_1\Gamma_1(X_0) = \Gamma(X_0) = R(X_0). \quad (3.3)$$

We are using the fact that X_0 is a minimizer for G_1 , though not uniquely so. The first inequality is strict unless $X = X_0$. Here we set $R(X) = \mathcal{E}_R(X)$, etc., for ease of notation. ♠

3.3 The General Approach

Now we consider 5-point configurations. We know that the TBP is a minimizer for Γ_0 and Γ_1 , though not uniquely so. Suppose that $\Gamma_2, \Gamma_3, \Gamma_4$ are 3 more functions. We have in mind the various triples which arise in the Forcing Lemma.

The distances involved in *TBP* are $\sqrt{2}, \sqrt{3}, \sqrt{4}$. (Writing $\sqrt{4}$ for 2 makes the equations look nicer.) Let R be some function defined on $(0, 2]$. Suppose we can find a combination

$$\Gamma = a_0\Gamma_0 + \dots + a_4\Gamma_4, \quad a_1, a_2, a_3, a_4 > 0. \quad (3.4)$$

such that

$$\Gamma(x) \leq R(x), \quad \Gamma(\sqrt{m}) = f(\sqrt{m}), \quad m = 2, 3, 4. \quad (3.5)$$

Suppose also that $\Gamma_j(X) > \Gamma_j(TBP)$ for $j = 2, 3, 4$. Then $R(X) > R(TBP)$. The proof is just like what we did for the tetrahedron:

$$R(X) \geq \Gamma(X) = \sum a_i \Gamma_i(X) > \sum a_i \Gamma_i(TBP) = \Gamma(TBP) = R(TBP). \quad (3.6)$$

Note that we have not given the minimal hypotheses needed to get this inequality to work.

Here is how we find the coefficients $\{a_i\}$. We impose the 5 conditions

- $\Gamma(x) = R(x)$ for $x = \sqrt{2}, \sqrt{3}, \sqrt{4}$.
- $\Gamma'(x) = R'(x)$ for $x = \sqrt{2}, \sqrt{3}$.

Here $R' = dR/dx$ and $\Gamma' = d\Gamma/dx$. These 5 conditions give us 5 linear equations in 5 unknowns. In the cases described below, the associated matrix is invertible and there is a unique solution. The proof of the Forcing Lemma, in each case, involves solving the matrix equation, checking positivity of the coefficients, and checking the under-approximation property. In this chapter we will present the solutions to the matrix equations, and show plots of the coefficients and sample plots of the under approximations.

3.4 Case 1

The triple is G_2, G_4, G_6 and $s \in (0, 6]$ and $R = R_s$, where $R_s(r) = r^{-s}$. We will suppress the constant s whenever it seems possible to do it without causing confusion. When we want to denote the dependence of Γ on the parameter, we write $\Gamma_{(s)}$. That is, $\Gamma_{(s)} = \sum a_i(s) \Gamma_i$. We need to find the coefficients a_0, a_1, a_2, a_3, a_4 , and we will also keep track of the quantity

$$\delta = 2\Gamma'(2) - 2R'(2). \quad (3.7)$$

The solution is given by

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ \delta \end{bmatrix} = \frac{1}{792} \begin{bmatrix} 0 & 792 & 1152 & -1944 & -54 & -288 & 0 \\ 792 & -1254 & -96 & 1350 & 87 & 376 & 0 \\ 528 & -312 & -216 & -39 & -98 & 0 & 0 \\ -66 & 48 & 18 & 6 & 10 & 0 & 0 \\ -6336 & -9216 & 15552 & 432 & 2304 & 792 & 0 \end{bmatrix} \begin{bmatrix} 2^{-s/2} \\ 3^{-s/2} \\ 4^{-s/2} \\ s2^{-s/2} \\ s3^{-s/2} \\ s4^{-s/2} \end{bmatrix}.$$

The left side of Figure 3.1 shows a graph of

$$80a_1, \quad 200a_2, \quad 2000a_3, \quad 10000a_4,$$

considered as functions of the exponent s . Here a_1, a_2, a_3, a_4 are colored darkest to lightest. The completely unimportant positive multipliers are present so that we get a nice picture. It turns out that a_3 goes negative between 6 and 6.1, so the interval $(0, 6]$ is fairly near to the maximal interval of positivity.

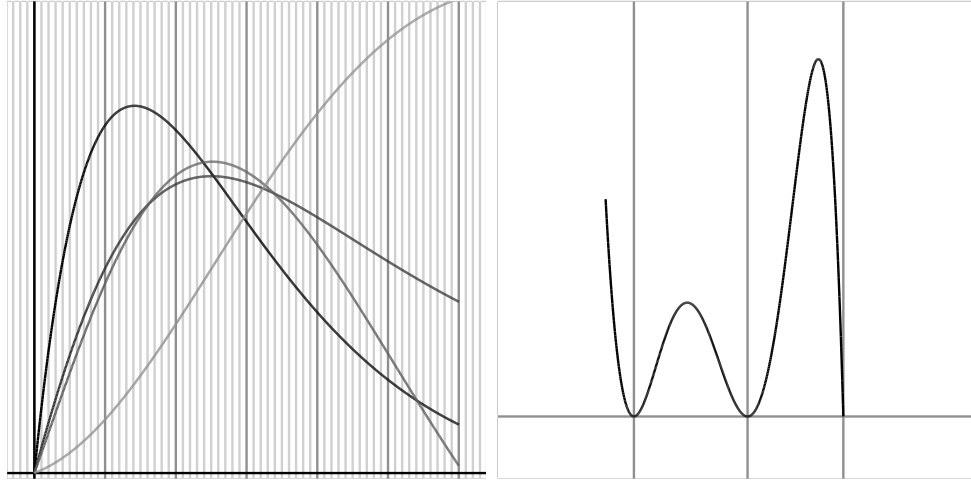


Figure 3.1: Plots for Case 1.

The right hand of Figure 3.1 shows some positive multiple of

$$H_3(r) = \left(1 - \frac{\Gamma_{(3)}(r)}{R_3(r)}\right)$$

plotted from $r = 4/3$ to $r = 2$. For $r < 4/3$ the graph is even more positive, but it is hard to fit the whole thing into a nice picture. The horizontal line is the x -axis and the vertical lines are $x = \sqrt{2}, \sqrt{3}, \sqrt{4}$. The fact that $H_3 \geq 0$ implies that $\Gamma_{(3)}(r) \leq R_3(r)$ for all $r \in (0, 2]$, as desired. The reader can use our computer program and get a much more interactive picture. In particular, you can adjust the endpoint of the plot, and see what happens for other values of s ,

Since $R_3 > 0$, the fact that $H_3 \geq 0$ implies that $\Gamma_{(3)}(r) \leq R_3(r)$ for all $r \in (0, 2]$. The right hand plot looks similar for other parameters $s \in (0, 6]$, as the reader can see from my program. Note that H_3 has negative slope at $r = 2$ and this corresponds to $\delta(3) > 0$.

3.5 Case 2

Here the triple is $G_2, G_5, G_{10}^{\#\#}$. We have

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ \hat{a}_4 \\ \delta \end{bmatrix} = \frac{1}{268536} \begin{bmatrix} 0 & 0 & 268536 & 0 & 0 & 0 \\ 88440 & 503040 & -591480 & -4254 & -65728 & 0 \\ -77586 & -249648 & 327234 & 2361 & 65896 & 0 \\ 41808 & -19440 & -22368 & -2430 & -9076 & 0 \\ -402 & 264 & 138 & 33 & 68 & 0 \\ -707520 & -4024320 & 4731840 & 34032 & 525824 & 268536 \end{bmatrix} \begin{bmatrix} 2^{-s/2} \\ 3^{-s/2} \\ 4^{-s/2} \\ s2^{-s/2} \\ s3^{-s/2}, \\ s4^{-s/2} \end{bmatrix}.$$

Figure 3.2 does for Case 2 what Figure 3.1 does for Case 1. This time the left hand side plots

$$500a_1 \quad 500a_2, \quad 5000a_3, \quad 500000a_4.$$

for $s \in [6, 13]$. The coefficients a_1, a_2, a_3 go negative for s just a tiny bit larger than 13. I worked hard to find the function $\Gamma_4 = G_{10} + 28G_5 + 102G_2$ so that we could get all the way up to $s = 13$. The right hand side shows a plot of H_{10} from $r = 5/4$ to $r = 2$. Since H_{10} has negative slope at $r = 2$ we have $\delta(10) > 0$.

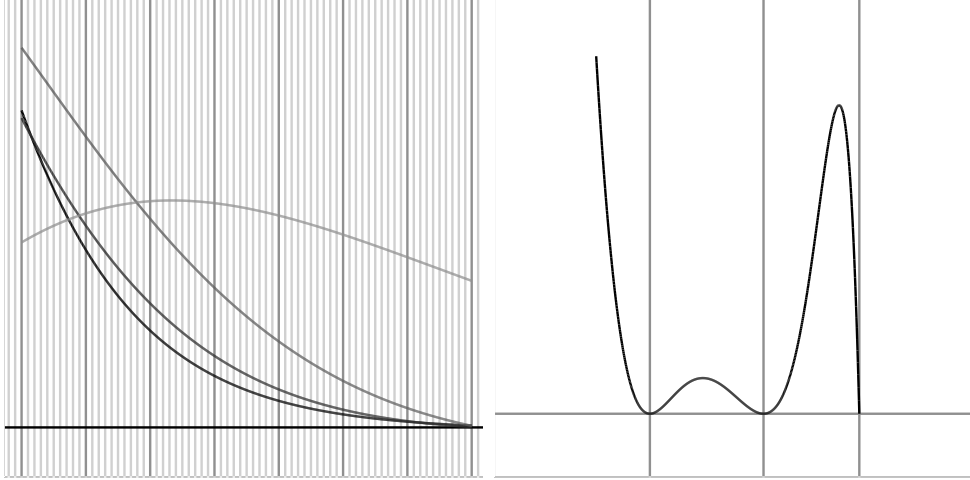


Figure 3.2: Plot for Case 2.

The interested reader can use our program to see what these coefficients look like for parameter value $s < 6$.

3.6 Case 3

Here the triple is $G_2, G_5^\flat, G_{10}^\sharp$. We have

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ \hat{a}_4 \\ \delta \end{bmatrix} = \frac{1}{268536} \begin{bmatrix} 0 & 0 & 268536 & 0 & 0 & 0 \\ 947112 & 131520 & -1078632 & -50694 & -259072 & 0 \\ -91254 & -240672 & 331926 & 3483 & 68208 & 0 \\ 35778 & -15480 & -20298 & -1935 & -8056 & 0 \\ -402 & 264 & 138 & 33 & 68 & 0 \\ 174268608 & 24199680 & -198468288 & -9327696 & -47669248 & 268536 \end{bmatrix}$$

This matrix is quite similar to the one in the previous case, because we are essentially still taking combinations of $G_0, G_1, G_2, G_5, G_{10}$. We are just grouping the functions differently. Figure 3.3 does for Case 3 what Figure 3.2 does for Case 2. This time we plot

$$500a_1 \quad 15000a_2, \quad 20000a_3, \quad 500000a_4,$$

for $s \in [13, 16]$. The coefficients a_1, a_2, a_3 go negative for s just a tiny bit larger than 15.05. In particular, everything up to and including our cutoff of $5 + 25/512$ is covered. The right hand side shows a plot of H_{14} from $r = 5/4$ to $r = 2$. Since H_{14} has negative slope at $r = 2$ we have $\delta(14) > 0$.

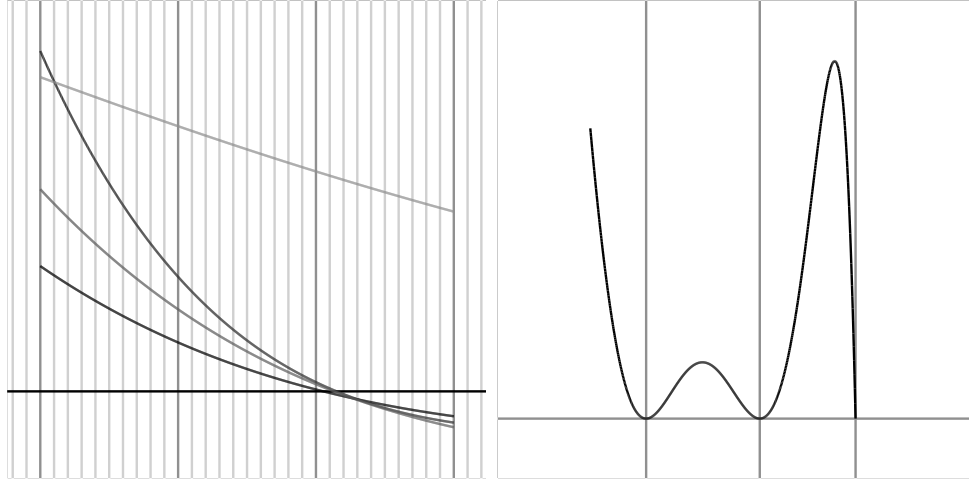


Figure 3.3: Plot for Case 3.

The interested reader can use our program to see what these coefficients look like for parameter value $s < 13$.

3.7 Case 4

Now we consider the negative case. The triple is (G_2, G_3, G_5) . The solution here is

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ \delta \end{bmatrix} = \frac{1}{144} \begin{bmatrix} 0 & 0 & -144 & 0 & 0 & 0 \\ -312 & -96 & 408 & 24 & 80 & 0 \\ 684 & -288 & -396 & -54 & -144 & 0 \\ -402 & 264 & 138 & 33 & 68 & 0 \\ 30 & -24 & -6 & -3 & -4 & 0 \\ 2496 & 768 & -3264 & -192 & -640 & -144 \end{bmatrix} \begin{bmatrix} 2^{-s/2} \\ 3^{-s/2} \\ 4^{-s/2} \\ s2^{-s/2} \\ s3^{-s/2} \\ s4^{-s/2} \end{bmatrix}.$$

The left side of Figure 3.4 shows a graph of

$$2a_1, \quad 100a_2, \quad 1000a_3, \quad 10000a_4,$$

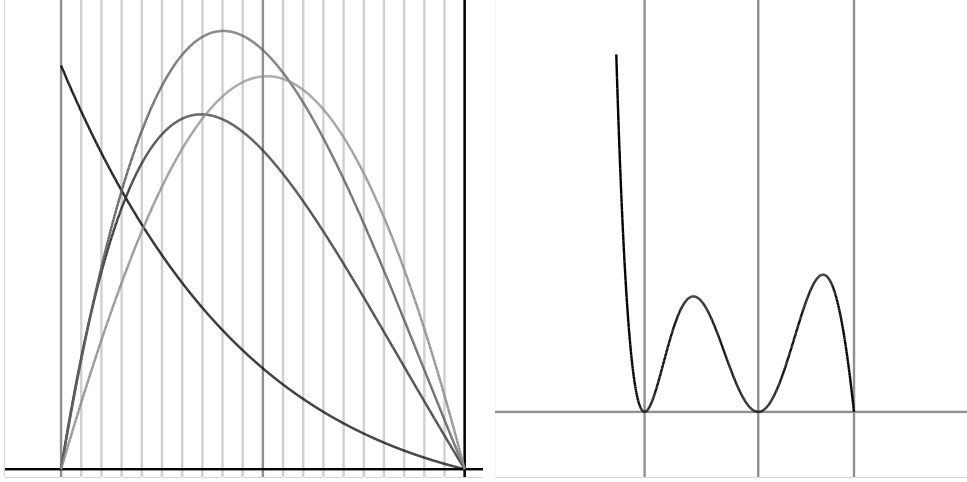


Figure 3.4: Plots for Case 4.

The right side of Figure 3.1 shows a plot of some constant times

$$H_{-1} = \left(\frac{\Gamma_{(-1)}(r)}{R_{-1}(r)} - 1 \right) \quad (3.8)$$

Note that H_{-1} has negative slope at $r = 2$, and this corresponds to $\delta(-1) > 0$.

3.8 Have Fun Playing Around

I encourage you to have fun playing around with my computer program. The program allows you to specify 5 functions of the form

$$AG_a + BG_b + CG_c,$$

and then it will show the above kinds of plots. Here A, B, C are integers and a, b, c are positive integers. The program also does a rough and ready calculation to decide where TBP has lower energy than any FP, so you can adjust the combinations so as to get things which would be useful for proving that the the TBP is a minimizer for power laws within a certain exponent range.

My program will also carry out rigorous positivity proofs along the lines of what we explain in the next three chapters. In this way you can cook up your own versions of the Forcing Lemma. For instance, the triple

$$\Gamma_2, \Gamma_4, \Gamma_8 + 4\Gamma_4 + 32\Gamma_2$$

is forcing on $[0, 11]$. I found all these crazy combinations using my program.

Chapter 4

Polynomial Approximations

This is the first of two chapters which develop machinery to prove the positivity claims needed for the Forcing Lemma – i.e. that the functions which appear to be positive (or non-negative) from the plots in the previous chapter really are positive (or non-negative). The functions involved in the Forcing Lemma are not polynomials. Thus, in this chapter we discuss a method for approximating these functions with polynomials. Basically, the method combines Taylor’s Theorem with the notion of an *interval polynomial*.

4.1 Rational Intervals

We define a *rational interval* to be an interval of the form $I = [L, R]$ where $L, R \in \mathbf{Q}$ and $L \leq R$. For each operation $*$ $\in \{+, -, \times\}$ we define

$$I_1 * I_2 = [\min(S), \max(S)], \quad S = \{L_1 * L_2, L_1 * R_2, R_1 * L_2, R_1 * R_2\}. \quad (4.1)$$

These operations are commutative. The definition is such that $r_j \in I_j$ for $j = 1, 2$ then $r_1 * r_2 \in I_1 * I_2$. Moreover, $I_1 * I_2$ is the minimal interval with this property. The minimality property implies that our laws are both associative and distributive:

- $(I_1 + I_2) \pm I_3 = I_1 + (I_2 \pm I_3)$.
- $I_1 \times (I_2 \pm I_3) = (I_1 \times I_2) \pm (I_1 \times I_3)$.

We also can raise a rational interval to a nonnegative integer power:

$$I^k = I \times \dots \times I \quad \text{k times.} \quad (4.2)$$

With the operations above, the set of rational intervals forms a *commutative semi-ring*. The 0 element is $[0, 0]$ and the 1-element is $[1, 1]$. Additive inverses do not make sense, and so we don't get a ring. Even though the semi-ring structure is nice and even beautiful, we technically do not need it. The main thing the semi-ring structure does is allow us to write complicated algebraic expressions without specifying the exact order of operations needed to evaluate the expressions. Were some of these laws to fail we would just pick some order of operations and this would suffice for our proofs. Indeed, when we perform interval arithmetic based on floating point operations we lose the semi-ring structure but still retain what we need for our proofs.

4.2 Interval Polynomials

An *interval polynomial* is an expression of the form

$$I_0 + I_1t + \dots + I_nt^n. \quad (4.3)$$

in which each coefficient is an interval and t is a variable meant to be taken in $[0, 1]$. Given the rules above, interval polynomials may be added, subtracted or multiplied, in the obvious way. The set of interval polynomials again forms a commutative semi-ring.

We think of an ordinary polynomial as an interval polynomial, *via* the identification

$$a_0 + \dots + a_nt^n \implies [a_0, a_0] + \dots + [a_n, a_n]t^n. \quad (4.4)$$

(Formally we have an injective semi-ring morphism from the polynomial ring into the interval polynomial semi-ring.) We think of a constant as an interval polynomial of degree 0. Thus, if we have some expression which appears to involve constants, ordinary polynomials, and interval polynomials, we interpret everything in sight as an interval polynomial and then perform the arithmetic operations needed to simplify the expression.

Let \mathcal{P} be the above interval polynomial. We say that \mathcal{P} *traps* the ordinary polynomial

$$C_0 + C_1t + \dots + C_nt^n \quad (4.5)$$

of the same degree if $C_j \in I_j$ for all j . We define the *min* of an interval polynomial to be the polynomial whose coefficients are the left endpoints of the intervals. We define the *max* similarly. If \mathcal{P} is an interval polynomial which traps an ordinary polynomial, then

- \mathcal{P} traps \mathcal{P}_{\min} .
- \mathcal{P} traps \mathcal{P}_{\max} .
- For all $t \in [0, 1]$ we have $\mathcal{P}_{\min}(t) \leq P(t) \leq \mathcal{P}_{\max}(t)$.

Our arithmetic operations are such that if the polynomial \mathcal{P}_j traps the polynomial P_j for $j = 1, 2$, then $\mathcal{P}_1 * \mathcal{P}_2$ traps $P_1 * P_2$. Here $*$ $\in \{+, -, \times\}$.

4.3 A Bound on the Power Functions

We are going to use Taylor's Theorem with Remainder to approximate the power functions. As a preliminary step, we need some *priori* bound on the remainder term over a suitable interval. We work in the interval $[-2, 16]$, which covers all the parameters of interest to us and a bit more.

Lemma 4.3.1 (Remainder Bound)

$$\left| \frac{d^{12}}{ds^{12}} m^{-s/2} \right| < 1 \quad \forall s \in [-2, 16], \quad \forall m = 2, 3, 4.$$

Proof: Note that $m^{-s/2} \leq m \leq 4$ when $s \in [-2, 16]$. The max occurs when $s = -2$. Also, one can check that $\log(4) < \sqrt{2}$. Hence $\log(4)^{12} < 64$. Hence

$$\frac{d^{12}}{ds^{12}} m^{-s/2} = \frac{m^{-s/2}}{4096 \log(m)} < \frac{m^{-s/2}}{64} \leq \frac{4}{64} < 1.$$

♠

Remark: Note that we get the much better bound of $1/16$. Whenever possible, we will be generous with our bounds because we want our results to be robustly true.

4.4 Approximation of Power Combos

Suppose that $Y = (a_2, a_3, a_4, b_2, b_3, b_4)$ is a 6-tuple of rational numbers. The coefficients produced by the Forcing Lemma construction have the form

$$C_Y(x) = a_2 2^{-s/2} + a_3 3^{-s/2} + a_4 4^{-s/2} + b_2 s 2^{-s/2} + b_3 s 3^{-s/2} + b_4 s 4^{-s/2} \quad (4.6)$$

evaluated on the interval $[-2, 16]$. We call such expressions *power combos*.

For each even integer $2k = -2, \dots, 16$, we will construct rational polynomials $A_{Y,2k,-}$ and $A_{Y,2k,+}$ and $B_{Y,2k,-}$ and $B_{Y,2k,+}$ such that

$$A_{Y,2k,-}(t) \leq C_Y(2k - t) \leq A_{Y,2k,+}(t), \quad t \in [0, 1]. \quad (4.7)$$

$$B_{Y,2k,-}(t) \leq C_Y(2k + t) \leq B_{Y,2k,+}(t), \quad t \in [0, 1]. \quad (4.8)$$

We ignore the cases $(2, -)$ and $(16, +)$.

The basic idea is to use Taylor's Theorem with Remainder:

$$m^{-s/2} = \sum_{j=0}^{11} \frac{(-1)^j \log(m)^j}{m^k 2^j j!} (s - 2k)^j + \frac{E_s}{12!} (s - 2k)^{12}. \quad (4.9)$$

Here E_s is the 12th derivative of $m^{-s/2}$ evaluated at some point in the interval. Note that the only dependence on k is the term m^k in the denominator, and this is a rational number.

The difficulty with this approach is that the coefficients of the above Taylor series are not rational. We get around this trick by using interval polynomials. We first pick specific intervals which trap $\log(m)$ for $m = 2, 3, 4$. We choose the intervals

$$L_2 = \left[\frac{25469}{36744}, \frac{7050}{10171} \right], \quad L_3 = \left[\frac{5225}{4756}, \frac{708784}{645163} \right], \quad L_4 = \left[\frac{25469}{18372}, \frac{345197}{249007} \right].$$

Each of these intervals has width about 10^{-10} . I found them using Mathematica's Rationalize function. It is an easy exercise to check that $\log(m) \in L_m$ for $m = 2, 3, 4$. According to our Remainder Lemma, we always have $|E_s| < 1$ in the series expansion from Equation 4.9. Fixing k we introduce the interval Taylor series

$$A_m(t) = \sum_{j=0}^{11} \frac{(+1)^j (L_m)^j}{m^k 2^j j!} t^j + \left[-\frac{1}{12!}, \frac{1}{12!} \right] t^{12}. \quad (4.10)$$

$$B_m(t) = \sum_{j=0}^{11} \frac{(-1)^j (L_m)^j}{m^k 2^j j!} t^j + \left[-\frac{1}{12!}, \frac{1}{12!} \right] t^{12}. \quad (4.11)$$

By construction A_m traps the Taylor series expansion from Equation 4.9 when it is evaluated at $t = s - 2k$ and $t \in [0, 1]$. Likewise B_m traps the

Taylor series expansion from Equation 4.9 when it is evaluated at $t = 2k - s$ and $t \in [0, 1]$. Define

$$\begin{aligned} A_Y(t) = & a_2 A_2(t) + a_3 A_3(t) + a_4 A_4(t) + \\ & b_2(2k - t)A_2(t) + b_3(2k - t)A_3(t) + b_4(2k - t)A_4(t) \end{aligned} \quad (4.12)$$

Remark: It took an embarrassing amount of trial and error before I realized that the coefficients $b_m(2k - t)$ are correct.

Likewise, define

$$\begin{aligned} B_Y(t) = & a_2 B_2(t) + a_3 B_3(t) + a_4 B_4(t) + \\ & b_2(2k + t)B_2(t) + b_3(2k + t)B_3(t) + b_4(2k + t)B_4(t) \end{aligned} \quad (4.13)$$

By construction, $A_Y(t)$ traps $C_Y(2k - t)$ when $t \in [0, 1]$ and $B_Y(t)$ traps $C_Y(2k + t)$ when $t \in [0, 1]$.

Finally, we define

$$\begin{aligned} A_{Y,2k,-} &= (A_Y)_{\min}, & A_{Y,2k,+} &= (A_Y)_{\max}, \\ B_{Y,2k,-} &= (B_Y)_{\min}, & B_{Y,2k,+} &= (B_Y)_{\max}. \end{aligned} \quad (4.14)$$

By construction these polynomials satisfy Equations 4.7 and 4.8 respectively for each $k = -1, \dots, 8$. These are our under and over approximations.

The reader can print out these approximations, either numerically or exactly, using my program. The reader can also run the debugger and see that these functions really do have the desired approximation properties.

Chapter 5

Positive Dominance

In the previous chapter we explained how to approximate the expressions that arise in the Forcing Lemma by polynomials. When we make these approximations, we reduce the Forcing Lemma to statements that various polynomials are positive (or non-negative) on various domains. The next chapter explains how this goes. In this chapter we explain how we prove such positivity results about polynomials. I will explain my Positive Dominance Algorithm. It is entirely possible to me that this algorithm is known to some community of mathematicians, but I thought of it myself and have not seen it in the literature.

5.1 The Weak Positive Dominance Criterion

Let

$$P(x) = a_0 + a_1x + \dots + a_nx^n \tag{5.1}$$

be a polynomial with real coefficients. Here we describe a method for showing that $P \geq 0$ on $[0, 1]$,

Define

$$A_k = a_0 + \dots + a_k. \tag{5.2}$$

We call P *weak positive dominant* (or *WPD* for short) if $A_k \geq 0$ for all k and $A_n > 0$.

Lemma 5.1.1 *If P is weak positive dominant, then $P > 0$ on $(0, 1]$.*

Proof: The proof goes by induction on the degree of P . The case $\deg(P) = 0$ follows from the fact that $a_0 = A_0 > 0$. Let $x \in (0, 1]$. We have

$$\begin{aligned} P(x) &= a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \geq \\ &= a_0x + a_1x + a_2x^2 + \cdots + a_nx^n = \\ &= x(A_1 + a_2x + a_3x^2 + \cdots + a_nx^{n-1}) = xQ(x) > 0 \end{aligned}$$

Here $Q(x)$ is weak positive dominant and has degree $n - 1$. ♠

We can augment this criterion by subdivision. Given $I = [a, b] \subset \mathbf{R}$, let A_I be one of the two affine maps which carries $[0, 1]$ to I . We call the pair (P, I) *weak positive dominant* if $P \circ A_I$ is WPD. If (P, I) is WPD then $P \geq 0$ on $(a, b]$, by Lemma 5.1.1. For instance, if P is WPD on $[0, 1/2]$ and $[1/2, 1]$ then $P > 0$ on $(0, 1)$.

5.2 The Positive Dominance Algorithm

As we suggested briefly above, the WPD *criterion* is really a step in a recursive subdivision algorithm. In explaining the algorithm, I will work with positive dominance rather than weak positive dominance, because the main application involves positive dominance. (For most of our proofs, the WPD criterion is enough.) I will treat the case when the domain is the unit cube. I have used this method extensively in other contexts. See e.g. [S2] and [S3]. In particular, in [S3] I explain the method for arbitrary polytopes.

Given a multi-index $I = (i_1, \dots, i_k) \in (\mathbf{N} \cup \{0\})^k$ we let

$$x^I = x_1^{i_1} \cdots x_k^{i_k}. \quad (5.3)$$

Any polynomial $F \in \mathbf{R}[x_1, \dots, x_k]$ can be written succinctly as

$$F = \sum a_I X^I, \quad a_I \in \mathbf{R}. \quad (5.4)$$

If $I' = (i'_1, \dots, i'_k)$ we write $I' \leq I$ if $i'_j \leq i_j$ for all $j = 1, \dots, k$. We call F *positive dominant* (PD) if

$$A_I := \sum_{I' \leq I} a_{I'} > 0 \quad \forall I, \quad (5.5)$$

Lemma 5.2.1 *If P is PD, then $P > 0$ on $[0, 1]^k$.*

Proof: When $k = 1$ the proof is the same as in Lemma 5.1.1, once we observe that also $P(0) > 0$. Now we prove the general case. Suppose the coefficients of P are $\{a_I\}$. We write

$$P = f_0 + f_1 x_k + \dots + f_m x_k^m, \quad f_j \in \mathbf{R}[x_1, \dots, x_{k-1}]. \quad (5.6)$$

Let $P_j = f_0 + \dots + f_j$. A typical coefficient in P_j has the form

$$b_J = \sum_{i=1}^j a_{Ji}, \quad (5.7)$$

where J is a multi-index of length $k-1$ and Ji is the multi-index of length k obtained by appending i to J . From equation 5.7 and the definition of PD, the fact that P is PD implies that P_j is PD for all j . ♠

The positive dominance criterion is not that useful in itself, but it feeds into a powerful divide-and-conquer algorithm. We define the maps

$$\begin{aligned} A_{j,1}(x_1, \dots, x_k) &= (x_1, \dots, x_{j-1} \frac{x_j + 0}{2}, x_{j+1}, \dots, x_k), \\ A_{j,2}(x_1, \dots, x_k) &= (x_1, \dots, x_{j-1} \frac{x_j + 1}{2}, x_{j+1}, \dots, x_k), \end{aligned} \quad (5.8)$$

We define the j th *subdivision* of P to be the set

$$\{P_{j1}, P_{j2}\} = \{P \circ A_{j,1}, P \circ A_{j,2}\}. \quad (5.9)$$

Lemma 5.2.2 *$P > 0$ on $[0, 1]^k$ if and only if $P_{j1} > 0$ and $P_{j2} > 0$ on $[0, 1]^k$.*

Proof: By symmetry, it suffices to take $j = 1$. Define

$$[0, 1]_1^k = [0, 1/2] \times [0, 1]^{k-1}, \quad [0, 1]_2^k = [1/2, 1] \times [0, 1]^{k-1}. \quad (5.10)$$

Note that

$$A_1([0, 1]^k) = [0, 1]_1^k, \quad B_1 \circ A_1([0, 1]^k) = [0, 1]_2^k. \quad (5.11)$$

Therefore, $P > 0$ on $[0, 1]_1^k$ if and only if $P_{j_1} > 0$ on $[0, 1]^k$. Likewise $P > 0$ on $[0, 1]_2^k$ if and only if $P_{j_2} > 0$ on $[0, 1]^k$. ♠

Say that a *marker* is a non-negative integer vector in \mathbf{R}^k . Say that the *youngest entry* in the the marker is the first minimum entry going from left to right. The *successor* of a marker is the marker obtained by adding one to the youngest entry. For instance, the successor of $(2, 2, 1, 1, 1)$ is $(2, 2, 2, 1, 1)$. Let μ_+ denote the successor of μ .

We say that a *marked polynomial* is a pair (P, μ) , where P is a polynomial and μ is a marker. Let j be the position of the youngest entry of μ . We define the *subdivision* of (P, μ) to be the pair

$$\{(P_{j_1}, \mu_+), (P_{j_2}, \mu_-)\}. \quad (5.12)$$

Geometrically, we are cutting the domain in half along the longest side, and using a particular rule to break ties when they occur. Now we have assembled the ingredients needed to explain the algorithm.

Divide-and-Conquer Algorithm:

1. Start with a list LIST of marked polynomials. Initially, LIST consists only of the marked polynomial $(P, (0, \dots, 0))$.
2. Let (Q, μ) be the last element of LIST. We delete (Q, μ) from LIST and test whether Q is positive dominant.
3. Suppose Q is positive dominant. We go back to Step 2 if LIST is not empty. Otherwise, we halt.
4. Suppose Q is not positive dominant. we append to LIST the two marked polynomials in the subdivision of (Q, μ) and then go to Step 2.

If the algorithm halts, it constitutes a proof that $P > 0$ on $[0, 1]^k$. Indeed, the algorithm halts if and only if $P > 0$ on $[0, 1]^k$.

Parallel Version: Here is a variant of the algorithm. Suppose we have a list $\{P_1, \dots, P_m\}$ of polynomials and we want to show that at least one of them is positive at each point of $[0, 1]^k$. We do the following

1. Start with m lists LIST(j) for $j = 1, \dots, m$ of marked polynomials. Initially, LIST(j) consists only of the marked polynomial $(P_j, (0, \dots, 0))$.
2. Let (Q_j, μ) be the last element of LIST(j). We delete (Q_j, μ) from LIST(j) and test whether Q_j is positive dominant. We do this for $j = 1, 2, \dots$ until we get a success or else reach the last index.
3. Suppose *at least one* Q_j is positive dominant. We go back to Step 2 if LIST(j) is not empty. (All lists have the same length.) Otherwise, we halt.
4. Suppose none of Q_1, \dots, Q_m is positive dominant. For each j we append to LIST(j) the two marked polynomials in the subdivision of (Q_j, μ) and then go to Step 2.

If this algorithm halts it constitutes a proof that at least one P_j is positive at each point of $[0, 1]^k$.

The Weak Version: We can run the positive dominance algorithm using the weak positive dominance criterion in place of the positive dominance criterion. If the algorithm runs to completion, it establishes that the given polynomial is non-negative on the unit cube, and almost everywhere positive.

5.3 Discussion

For polynomials in 1 variable, the method of Sturm sequences counts the roots of a polynomial in any given interval. An early version of this paper used Sturm sequences (to get more limited results) but I prefer the positive dominance criterion. The calculations for the positive dominance criterion are much simpler and easier to implement than Sturm sequences.

Another 1 dimensional technique that sometimes works is Descartes' Rule of Signs or (relatedly) Budan's Theorem. Both these techniques involve examining the coefficients and looking for certain patterns. In this way they are similar to positive dominance. However, they do not work as well. I could imagine placing Budan's Theorem inside an iterative algorithm, but I don't think it would work as well as the PDA.

There are generalizations of Sturm sequences to higher dimensions, and also other positivity criteria (such as the Handelman decomposition) but I bet they don't work as well as the positive dominance algorithm. Also, I don't see how to do the parallel positive dominance algorithm with these other methods.

The positive dominance algorithm works so well that one can ask why I didn't use it to deal with the calculations related to the Big and Small Theorems discussed in §2.2. For such calculations, the polynomials involved, when expanded out, have a vast number of terms. The calculation is not feasible. I briefly tried to set up the problem for the energy G_3 and already the number of terms was astronomical.

Chapter 6

Proof of The Forcing Lemma

6.1 Positivity of the Coefficients

Case 1: We consider a_1 on $(0, 6]$ in detail.

- We set

$$Y = (792, 1152, -1944, -54, -288, 0),$$

the row of the relevant matrix corresponding to a_1 . By construction, $a_1(x) = C_Y(x)$, the power combo defined in §4.4.

- We verify that the 6 under-approximations

$$A_{Y,0,+}, A_{Y,2,-}, A_{Y,2,+}, A_{Y,4,-}, A_{4,+}, A_{6,-}$$

are either WPD on $[0, 1]$ or WPD on each of $[0, 1/2]$ and $[1/2, 1]$ in all cases. Hence these functions are positive on $(0, 1]$. See §5.1.

- Since $A_{Y,k,+}(t) \leq a_1(k+t)$ for $t \in [0, 1]$ we see that $a_1 > 0$ on $(k, k+1]$ for $k = 0, 2, 4$.
- Since $A_{Y,k,-}(t) \leq a_1(k-t)$ for $t \in [0, 1]$ we see that $a_1 > 0$ on $[k, k+1)$ for $k = 1, 3, 5$.
- We check by direct inspection that $a_1(x) > 0$ at the values $x = 1, 2, 3, 4, 5, 6$.

The same argument works for a_2, a_3, a_4, δ on $(0, 6]$. We conclude that $a_1, a_2, a_3, a_4, \delta > 0$ on $(0, 6]$. By construction $a_0(s) = 2^{-s} > 0$. So, all coefficients are positive on $(0, 6]$. The statement $\delta > 0$ means $\Gamma'(2) > R'(2)$ for all $s \in (0, 6]$.

Case 2: We do the same thing on the interval $[6, 13]$, using the intervals $[6, 7], \dots, [12, 13]$. Again, every polynomial in sight is WPD on $[0, 1]$, and in fact PD on $[0, 1]$. Since we just check the WPD condition, we also check that our functions are positive at the integer values in $[6, 13]$ by hand. We conclude that $a_1, a_2, a_3, a_4, \delta > 0$ on $[6, 13]$.

Case 3: This case is different, because we are checking positivity on an interval whose right endpoint is not an endpoint. In all cases, we use the positive dominance algorithm with the subdivision variant that causes the intervals to be treated in order. The algorithm fails on the interval $[13, 16]$ but for each coefficient it passes on $[13, 14]$ and $[14, 15]$ and then we check the following:

1. The polynomial under-approximating a_1 is also PD on $[15, 481/32]$ and $[481/32, 963/64]$ and $[963/64, 3853/256]$.
2. The polynomial under-approximating a_2 is PD on $[15, 121/8]$.
3. The polynomial under-approximating a_3 is PD on $[15, 121/8]$.
4. The polynomial under-approximating a_4 is PD on $[15, 16]$.
5. The polynomial under-approximating δ is also PD on $[15, 481/32]$ and $[481/32, 963/64]$ and $[963/64, 1721/128]$.

The numbers

$$3853/256, \quad 121/8, \quad 1971/128$$

all (barely) exceed 15.05. Hence $a_1, a_2, a_3, a_4, \delta > 0$ on $[13, 15.05]$.

Case 4: This case is just like Case 1, except that here our function is $R_s(r) = -r^{-s}$. This means that $a_0(s) = -2^{-s} < 0$. The same argument as in Case 1 shows that $a_1, a_2, a_3, a_4, \delta > 0$ on $(-2, 0)$. We also check, using the same method that

$$\Gamma_{(s)}(0) = a_0 + 4a_1 + 16a_2 + 256a_3 + 1024a_5 < 0, \quad \forall s \in (-2, 0).$$

6.2 Under Approximation: Case 1

We have $s \in (0, 6]$. as we can. In particular, we set $R = R_s$, etc. We know already that $\Gamma(r) > 0$ and $R(r) > 0$ for all $r \in (0, 6]$. Define

$$H(r) = 1 - \frac{\Gamma(r)}{R(r)} = 1 - r^s \Gamma. \quad (6.1)$$

We just have to show (for each value of s) that $H \geq 0$ on $(0, 2)$. Let

$$H' = \frac{dH}{dr}. \quad (6.2)$$

Lemma 6.2.1 *H' has 4 simple roots in $(0, 2)$.*

Proof: We count roots with multiplicity. We have

$$H'(r) = -r^{s-1}(s\Gamma(r) + r\Gamma'(r)). \quad (6.3)$$

Combining Equation 6.3 with the general equation

$$rG'_k(r) = 2kG_k(r) - 8kG_{k-1}(r), \quad (6.4)$$

we see that the positive roots of $H'(r)$ are the same as the positive roots of

$$s\Gamma(r) + r\Gamma'(r) = (12 + s)a_4G_6(r) - 48a_4G_5(r) + \sum_{k=1}^4 b_k G_k(r) + b_0. \quad (6.5)$$

Here b_0, \dots, b_4 are coefficients we don't care about. Making the substitution $t = 4 - r^2$ and dividing through to make the polynomial monic, we see that the roots of H' in $(0, 2)$ are in bijection with the roots in $(0, 4)$ of

$$\psi(t) = t^6 - \frac{48}{12 + s}t^5 + c_4t^4 + c_3t^3 + c_2t^2 + c_1t + c_0. \quad (6.6)$$

Here c_0, \dots, c_4 are coefficients we don't care about. The change of coordinates is a diffeomorphism from $(0, 4)$ to $(0, 2)$ and so it carries simple roots to simple roots.

We just have to show that ψ has 4 simple roots in $(0, 4)$. Note that the sum of the 6 roots of ψ is $48/(12 + s) < 4$. This works because $s > 0$ here. The 4 roots of ψ we already know about are 1 and 2 and some number in

$(0, 1)$ and some number in $(1, 2)$. The sum of these roots exceeds 4 and so the remaining two roots cannot also be positive. Hence ψ has at most 5 roots in $(0, 4)$.

To finish the proof, it suffices to show that ψ has an even number of roots in $(0, 4)$. Let's call a function f *essentially positive from the left* at the point x_0 if there is an infinite sequence of values converging to x_0 from the left at which f is negative. We make the same definitions with *right* in place of *left* and *negative* in place of *positive*.

To prove our claim about the parity of the number of roots, it suffices to show that ψ is essentially negative from the right at 0 and essentially negative from the left at 4. For this purpose, it suffices to show that H' is essentially negative at 0 and negative at 2.

Note that $H(\epsilon) \rightarrow 1$ as $\epsilon \rightarrow 0$ because $R \rightarrow \infty$ and Γ is bounded. Also, both Γ and R are both positive. From these properties, we see that H' is essentially negative from the right at 0. At the same time, we know that $\Gamma(2) = R(2)$ and $\Gamma'(2) > R'(2)$. But then

$$H'(2) = \frac{\Gamma(2)R'(2) - \Gamma'(2)R(2)}{R^2(2)} = \frac{R'(2) - \Gamma'(2)}{R(2)} < 0. \quad (6.7)$$

This completes the proof that ψ has an even number of roots in $(0, 4)$. Since there are at most 5 such roots, and at least 4 distinct roots, there are exactly 4 roots. Since they are all distinct, they are all simple. ♠

How we mention the explicit dependence on s and remember that we are taking about H_s .

Lemma 6.2.2 $H_s''(\sqrt{2}) > 0$ and $H_s''(\sqrt{3}) > 0$ for all $s \in (0, 6]$.

Proof: We check directly that

$$H_3''(\sqrt{2}) > 0, \quad H_3''(\sqrt{3}) > 0.$$

It cannot happen that $H_s''(\sqrt{2}) = 0$ for other $s \in (0, 6]$ because then H'_s has a double root in $(0, 2)$. Hence $H_s''(\sqrt{2}) > 0$ for all $s \in (0, 6]$. The same argument shows that $H_s''(\sqrt{3}) > 0$ for all $s \in (0, 6]$. ♠

Now we set $H = H_s$ again.

Lemma 6.2.3 *For all sufficiently small $\epsilon > 0$ the quantities*

$$H(0 + \epsilon), \quad H(\sqrt{2} \pm \epsilon), \quad H(\sqrt{3} \pm \epsilon), \quad H(2 - \epsilon)$$

are positive.

Proof: We have seen already that $\lim_{\epsilon \rightarrow 0} H(\epsilon) = 1$. Likewise, we have seen that $H'(2) < 0$ and $H(2) = 0$. So $H(2 - \epsilon) > 0$ for all sufficiently small ϵ . Finally, the case of $\sqrt{2}$ and $\sqrt{3}$ follows from the previous lemma and the second derivative test. ♠

We have already proved that H' has exactly 4 simple roots in $(0, 2)$. In particular, the interval $(0, \sqrt{2})$ has no roots of H' and the intervals $(\sqrt{2}, \sqrt{3})$ and $(\sqrt{3}, 2)$ have 1 root each. Finally, we know that $H > 0$ sufficiently near the endpoints of all these intervals. If $H(x) < 0$ for some $x \in (0, 2)$, then x must be in one of the 3 intervals just mentioned, and this interval contains at least 2 roots of H' . This is a contradiction. This shows that $H \geq 0$ on $(0, 2)$ and thereby completes Case 1.

6.3 Under Approximation: Case 2

This time we have $s \in [6, 13]$ and everything is as in Case 1. All we have to do is show that the polynomial ψ , as Equation 6.6, has exactly 4 simple roots in $(0, 2)$. This time we have

$$\psi(t) = t^{10} - \frac{80}{20+s}t^9 + b_8t^8 + \dots + b_0. \quad (6.8)$$

Again these coefficients depend on s .

Remark: ψ only has 7 nonzero terms and hence can only have 6 positive real roots. The number of positive real roots is bounded above by Descartes' rule of signs. Unfortunately, ψ turns out to be alternating and so Descartes' rule of signs does not eliminate the case of 6 roots. This approach seems useless. The sum of the roots of ψ is less than 4, so it might seem as if we could proceed as in Case 2. Unfortunately, there are 10 such roots and this approach also seems useless. We will take another approach to proving what we want.

Lemma 6.3.1 *When $s = 6$ the polynomial ψ has 4 simple roots in $(0, 4)$.*

Proof: We compute explicitly that

$$\psi(t) = t^{10} - \frac{40}{13}t^9 + \frac{830304}{5785}t^5 - \frac{415152}{1157}t^4 + \frac{789255}{1157}t^2 - \frac{3264104}{5785}t + \frac{115060}{1157}.$$

This polynomial only has 4 real roots – the ones we know about. The remaining roots are all at least $1/2$ away from the interval $(0, 4)$ and so even a very crude analysis would show that these roots do not lie in $(0, 4)$. We omit the details.

Lemma 6.3.2 *Suppose, for all $s \in [6, 13]$ that ψ only has simple roots in $(0, 4)$. Then in all cases ψ has exactly 4 such roots.*

Proof: Let N_s denote the number of simple roots of ψ at the parameter s . The same argument as in Cases 1 and 2 shows that N_s is always even. Suppose s is not constant. Consider the infimal $u \in (6, 13]$ such that $N_u > 4$. The roots of ψ vary continuously with s . How could more roots move into $(0, 4)$ as $s \rightarrow u$?

One possibility is that such a root r_s approaches from the upper half plane or from the lower half plane. That is, r_s is not real for $s < u$. Since ψ is a real polynomial, the conjugate \bar{r}_s is also a root. The two roots r_s and \bar{r}_s are approaching $(0, 4)$ from either side. But then the limit

$$\lim_{s \rightarrow u} r_s$$

is a double root of ψ in $(0, 4)$. This is a contradiction.

The only other possibility is that the roots approach along the real line. Hence, there must be some $s < u$ such that both 0 and 4 are roots of ψ . But the same parity argument as in Case 1 shows $\psi(0) > 0$ for all $s \in [6, 13]$. ♠

To finish the proof we just have to show that ψ only has simple roots in $(0, 4)$ for all $s \in [6, 13]$. We bring the dependence on s back into our notation and write ψ_s . It suffices to show that that ψ_s and $\psi'_s = d\psi_s/dr$ do not simultaneously vanish on the rectangular domain $(s, r) \in [6, 13] \times [0, 4]$. This is a job for our method of positive dominance.

We will explain in detail what do on the smaller domain

$$(s, r) = [6, 7] \times [0, 4].$$

The proof works the same for the remaining 1×4 rectangles. The coefficients of ψ_s and ψ'_s are power combos in the sense of Equation 4.4.

We have rational vectors Y_0, \dots, Y_9 such that

$$\psi_s(r) = \sum_{j=0}^9 C_j r^j, \quad C_j = C_{Y_j}. \quad (6.9)$$

We have the under- and over-approximations:

$$A_j = A_{Y_j,6,+} \quad B_j = B_{Y_j,6,+} \quad (6.10)$$

We then define 2-variable under- and over-approximations:

$$\underline{\psi}(t, u) = \sum_{i=0}^9 A_i(t)(4u)^i, \quad \overline{\psi}(t, u) = \sum_{i=0}^9 B_i(t)(4u)^i. \quad (6.11)$$

We use $4u$ in these sums because we want our domains to be the unit square. We have

$$\underline{\psi}(t, u) \leq \psi_{6+t}(4u) \leq \overline{\psi}(t, u), \quad \forall (t, u) \in [0, 1]^2. \quad (6.12)$$

Now we do the same thing for ψ'_s . We have rational vectors $Y'_0, \dots, Y'_8 \in \mathcal{Q}^8$ which work for ψ' in place of ψ , and this gives under- and over-approximations $\underline{\psi}'$ and $\overline{\psi}'$ which satisfy the same kind of equation as Equation 6.12.

We run the parallel positive dominance algorithm on the set of functions $\{\underline{\phi}, -\overline{\phi}, \underline{\phi}', -\overline{\phi}'\}$ and the algorithm halts. This constitutes a proof that at least one of these functions is positive at each point. But then at least one of $\phi_s(r)$ or $\phi'_s(r)$ is nonzero for each $s \in [6, 13]$ and each $r \in [0, 4]$. Hence ψ_s only has simple roots in $(0, 4)$. This completes our proof in Case 2.

6.4 Under-Approximation: Case 3

This time we work in the interval $[13, 16]$. For both Case 3 and Case 4 we are finding some linear combination of the functions $G_0, G_1, G_2, G_5, G_{10}$ which matches the relevant power law at $\sqrt{2}, \sqrt{3}, \sqrt{4}$. Even though the coefficients and the functions are different in the two cases, the final linear combination is the same. Thus, we just have to re-do Case 3 on the intervals $[13, 14]$, $[14, 15]$, and $[15, 16]$. We do this, and it works out the same way. This completes our proof in Case 3.

6.5 Under Approximation: Case 4

This case is just like Case 1 except for some irritating sign changes. Mostly we will explain the differences between this case and Case 1. Here we show that $\Gamma_{(s)}(r) \leq R_s(r)$ for all $r \in (0, 2]$ and all $s \in (-2, 0)$. Here we have $R_s < 0$ so we want show that $\Gamma_{(s)}$ is even more negative. This time we define

$$H = \frac{\Gamma}{R} - 1. \quad (6.13)$$

We want to show that $H \geq 0$ on $(0, 2)$.

As in Case 1, we first prove that H' has exactly 4 simple roots in $(0, 2)$. We take the same steps as in Case 1, and this leads to the polynomial

$$\psi(t) = t^5 - \frac{40}{10+s}t^4 + b_3t^3 + b_2t^2 + b_1t + b_0. \quad (6.14)$$

The polynomial ψ has 5 roots counting multiplicities, and also 4 distinct roots we know about in $(0, 4)$. We just have to show that ψ has an even number of roots. As in Case 1, it suffices to show that H' is essentially negative from the right at 0 and $H'(2) < 0$. The calculation that $H'(2) < 0$ is the same as in Case 1. Since $R < 0$ on $(0, 2)$ and $R(0) = 0$ and $\Gamma(0) < 0$ we see that $H(r) \rightarrow \infty$ as $r \rightarrow 0$. But this shows that H' is essentially negative from the right at 0. This gives us our 4 simple roots at every parameter $s \in (-2, 0)$.

We make explicit checks that

$$H''_{(-1)}(\sqrt{2}) > 0, \quad H''_{(-1)}(\sqrt{3}) > 0.$$

Given this information, the rest of the proof is exactly like Case 1.

Part III

Divide and Conquer

Chapter 7

The Configuration Space

7.1 Normalized Configurations

The next 5 chapters are devoted to proving the Big Theorem and the Small Theorem from §2.2. In this chapter we study the moduli space in which the calculations take place.

We consider 5-point configurations $\hat{p}_0, \dots, \hat{p}_4 \in S^2$ with $\hat{p}_4 = (0, 0, 1)$. For $k = 1, 2, 3, 4$ let $p_k = \Sigma(\hat{p}_k)$ where Σ is the stereographic projection from Equation 1.4. Before we go further, we mention that inverse stereographic projection is given by

$$\Sigma^{-1}(x, y) = \left(\frac{2x}{1 + x^2 + y^2}, \frac{2y}{1 + x^2 + y^2}, 1 - \frac{2}{1 + x^2 + y^2} \right). \quad (7.1)$$

This is the inverse of the map Σ in Equation 1.4.

Setting $p_k = (p_{k1}, p_{k2})$, we normalize so that

- p_0 lies in the positive x -axis. That is, $p_{01} > 0$ and $p_{02} = 0$.
- $\|p_0\| \geq \max(\|p_1\|, \|p_2\|, \|p_3\|)$.
- $p_{12} \leq p_{22} \leq p_{32}$ and $0 \leq p_{22}$.

If, in addition, we are working with a monotone decreasing energy function, we add the condition that

$$p_{12} \leq 0. \quad (7.2)$$

If this condition fails then our normalization implies that all the points lie in the same hemisphere. But then one can decrease the energy by reflecting

one of the points across the hemisphere boundary. This argument can fail for the one energy function we consider which is not monotone, namely G_5^b .

We will sometimes augment these conditions in some circumstances, but the above is the baseline. We call such configurations *normalized*.

7.2 Compactness Results

Lemma 7.2.1 *Let Γ be any of $G_3, G_4, G_5^b, G_5, G_6, G_{10}^\sharp, G_{10}^{\sharp\sharp}$. Any normalized minimizer w.r.t. Γ has $p_0 \in [0, 4]$ and $\|p_k\| \leq 3/2$ for $k = 1, 2, 3$.*

Proof: The TBP has 6 bonds of length $\sqrt{2}$, and 3 bonds of length $\sqrt{3}$, and one bond of length 2. Hence,

$$G_k(TBP) = 3(2^{k+1} + 1), \quad k = 1, 2, 3, \dots \quad (7.3)$$

Suppose first that Γ is one of the energies above, but not G_3 or G_5^b . These two require special consideration. If $\|p_0\| \geq 4$ then the distance from \hat{p}_0 to $(0, 0, 1)$ is at most $d = \sqrt{4/17}$. We check this by computing $\Sigma^{-1}(4, 0)$, which would be the farthest point from $(0, 0, 1)$ with the given constraints. We check by direct calculation that $\Gamma(d) > \Gamma(TBP)$ in all cases. This single bond contributes too much to the energy all by itself.

If the second condition fails then we have $\|p_k\| > 3/2$ for some k and also $\|p_0\| > 3/2$. The distance from \hat{p}_k and \hat{p}_0 from \hat{p}_4 is at most $d' = 4/\sqrt{13}$. In all cases we compute that $2\Gamma(d') > \Gamma(TBP)$. For example $G_4(TBP) = 99$ whereas $2\Gamma(d') \approx 117.616$. This deals with all cases except G_3 and G_5^b .

Now we deal with G_3 . We have $G_3(d) > G_3(TBP)$. This gives $\|p_0\| \leq 4$ as in the preceding case. However, now we see that $2G_3(d') = 42.4725$ whereas $G_3(TBP) = 51$. We will make up the rest of the energy by looking at the bonds which do not involve \hat{p}_4 .

The distance between any two points for the 4-point minimizer is at least 1, because otherwise the energy of the single bond, $3^3 = 27$, would exceed the energy $6 \times (4/3)^4 \in (14, 15)$ of the regular tetrahedron. Now observe that the function G_3 is convex decreasing on the interval $[1, 2]$. We can replace G_3 by a function G'_3 which is convex decreasing on $(0, 2]$ and agrees with G_3 on $[1, 2]$. For our minimizer, we have

$$\sum_{i < j < 4} G_3(\|\hat{p}_i - \hat{p}_j\|) = \sum_{i < j < 4} G'_3(\|\hat{p}_i - \hat{p}_j\|) \geq^* 6G'_3(\sqrt{8/3}) > 14.$$

The starred equality comes from the fact that the regular tetrahedron is the global minimizer, amongst 4-point configurations, for any convex decreasing potential. In conclusion, the sum of energies in the bonds not involving \widehat{p}_4 is at least 14. Since $14 + 42 > 51$, our configuration could not be a minimizer if we have $\|p_k\| > 3/2$ for some $k = 1, 2, 3$.

Finally, we deal with G_5^b . The preceding arguments used the fact that $\Gamma \geq 0$ on $(0, 2]$. We don't have this property for G_5^b but it suffices to prove the result for $\Gamma = G_5^b + 30$. A bit of calculus shows that $\Gamma > 0$ on $(0, 2]$. The preceding arguments seemed to use the fact that Γ is decreasing, and this is not the case in the present situation. However, all we really used is that Γ is decreasing on the intervals $(0, d]$ and $(0, d']$. A bit of calculus shows that our Γ here is decreasing on $(0, 1]$ and we have $d, d' < 1$.

We compute

$$\Gamma(TBP) = G_5(TBP) - 25G_1(TBP) + 300 = 120.$$

We also compute that $\Gamma(d) > 692 > \Gamma(TBP)$ and $2\Gamma(d') > 716 > \Gamma(TBP)$. The rest of the proof is the same as in the previous cases. ♠

Now we give a lengthy discussion which contributes nothing to the proof but sheds some light on what we are doing. The precise bounds given above are not really vital to our proof, though they do turn out to be very convenient. All we would really need, to get our calculations off the ground, is some (reasonable) constant C such that $\|p_0\| < C$ for any normalized configuration. Given such a bound, we could reconfigure our calculations (and estimates) to handle a wider, but still compact, domain of configurations. The point I am trying to make is that our approach is more robust than Lemma 7.2.1 might indicate.

Now let me turn the discussion to the function G_2 , which plays a special role in our proof because it appears in all cases of the Forcing Lemma. Fortunately, Tumanov has a very nice proof that the TBP is the minimizer w.r.t. G_2 , but one might wonder what we would have done without Tumanov's result. When we simply confine our points to the domain given by the lemmas above, our calculations run to completion for G_2 , and the local analysis of the Hessian turns out to work fine as well. Again, if we had any reasonable bound on $\|p_0\|$ we could get the Big Theorem to work for G_2 .

The arguments above do not work for G_2 , because $G_2(TBP) = 27$, whereas a single bond contributes at most 16 to the energy and the reg-

ular tetrahedron has G_2 energy $32/3$. Since $16 + 32/3 = 26 + \frac{2}{3} < 27$, we cannot use the argument above to get a bound. One way to scramble around this problem is to observe that the argument fails only if two points are very close together and either one makes something very close to a regular tetrahedron with the other 3. An examination of the equations in §3.2 would eventually yield such a result. In this case, an examination of more of the bonds would reveal the energy to be more than 27, and we would get our desired bound. Fortunately, we don't have to wander down this path.

7.3 The TBP Configurations

The TBP has two kinds of points, the two at the poles and the three at the equator. When ∞ is a polar point, the points p_0, p_1, p_2, p_3 are, after suitable permutation,

$$(1, 0), \quad (-1/2, -\sqrt{3}/2), \quad (0, 0), \quad (-1/2, +\sqrt{3}/2). \quad (7.4)$$

We call this the *polar configuration*. When ∞ is an equatorial point, the points p_0, p_1, p_2, p_3 are, after suitable permutation,

$$(1, 0), \quad (0, -1/\sqrt{3}), \quad (-1, 0), \quad (0, 1/\sqrt{3}). \quad (7.5)$$

We call this the *equatorial configuration*. We can visualize the two configurations together in relation to the regular 6-sided star. The black points are part of the polar configuration and the white points are part of the equatorial configuration. The grey point belongs to both configurations. The points represented by little squares are polar and the points represented by little disks are equatorial. The beautiful pattern made by these two configurations is not part of our proof, but it is nice to contemplate.

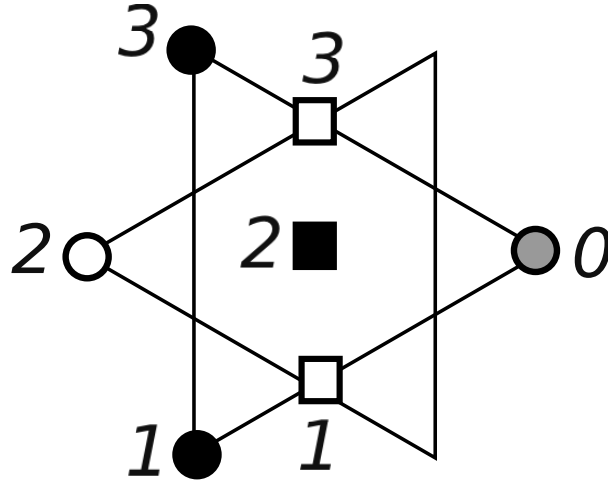


Figure 7.1: Polar and equatorial versions of the TBP.

We prefer ¹ the equatorial configuration, and we will use a trick to avoid ever having to search very near the polar configuration.

For each $k = 0, \dots, 4$ we introduce the quantity

$$\delta_k = \max_j \|\hat{p}_j - \hat{p}_k\|^2. \quad (7.6)$$

We square the distance just to get a rational function. We say that a normalized configuration is *totally normalized* if

$$\delta_4 \leq \delta_j, \quad j = 0, 1, 2, 3. \quad (7.7)$$

This condition is saying that the points in the configuration are bunched up around $(0, 0, 1)$ as much as possible. In particular, the polar TBP is not totally normalized and the equatorial TBP is totally normalized. By relabeling our points we can assume that our configurations are totally normalized.

7.4 Dyadic Blocks

For the moment we find convenient to only require that $p_k \in [-2, 2]^2$ for $k = 1, 2, 3$. Later on, we will enforce the stronger condition given by the

¹In an earlier version of this paper, which just proved the Big Theorem, I preferred the polar TBP and used a trick to avoid searching near the equatorial TBP. On further reflection, I found it better to switch my preferences. Either approach works for the Big Theorem, but for the Small Theorem it is much better to prefer the equatorial TBP.

lemmas above. Define

$$\square = [0, 4] \times [-2, 2]^2 \times [-2, 2]^2 \times [-2, 2]^2. \quad (7.8)$$

Any minimizer of any of the energies we consider is isometric to one which is represented by a point in this cube. This cube is our universe.

In 1 dimension, the *dyadic subdivision* of a line segment is the union of the two segments obtained by cutting it in half. In 2 dimensions, the *dyadic subdivision* of a square is the union of the 4 quarters that result in cutting the square in half along both directions. See Figure 7.2.



Figure 7.2: Dyadic subdivision

We say that a *dyadic segment* is any segment obtained from $[0, 4]$ by applying dyadic subdivision recursively. We say that a *dyadic square* is any square obtained from $[-2, 2]^2$ by applying dyadic subdivision recursively. We count $[0, 4]$ as a dyadic segment and $[-2, 2]^2$ as a dyadic square.

Hat and Hull Notation: Since we are going to be switching back and forth between the picture on the sphere and the picture in \mathbf{R}^2 , we want to be clear about when we are talking about solid bodies, so to speak, and when we are talking about finite sets of points. We let $\langle X \rangle$ denote the convex hull of any Euclidean subset. Thus, we think of a dyadic square Q as the set of its 4 vertices and we think of $\langle Q \rangle$ as the solid square having Q as its vertex set. Combining this with our notation for stereographic projection, we get the following notation, which we will use repeatedly.

- \widehat{Q} is a set of 4 co-circular points on S^2 .
- $\langle \widehat{Q} \rangle$ is a convex quadrilateral whose vertices are \widehat{Q} .
- $\widehat{\langle Q \rangle}$ is a “spherical patch” on S^2 , bounded by 4 circular arcs.

Good Squares: A dyadic square is *good* if it is contained in $[-3/2, 3/2]^2$ and has side length at most $1/2$. Note that a good dyadic square cannot cross the coordinate axes. The only dyadic square which crosses the coordinate

axes is $[-2, 2]^2$, and this square is not good. Our computer program will only do spherical geometry calculations on good squares.

Dyadic Blocks: We define a *dyadic block* to be a 4-tuple (Q_0, Q_1, Q_2, Q_3) , where Q_0 is a dyadic segment and Q_i is a dyadic square for $j = 1, 2, 3$. We say that a block is *good* if each of its 3 component squares is good. By Lemma 7.2.1, any energy minimizer for G_k is contained in a good block. Our algorithm in §10 quickly chops up the blocks in \square so that only good ones are considered.

The product

$$\langle B \rangle = \langle Q_0 \rangle \times \langle Q_1 \rangle \times \langle Q_2 \rangle \times \langle Q_3 \rangle \quad (7.9)$$

is a rectangular solid in the configuration space \square . On the other hand, the product

$$B = Q_0 \times Q_1 \times Q_2 \times Q_3 \quad (7.10)$$

is the collection of 128 vertices of $\langle B \rangle$. We call these the *vertex configurations* of the block.

Definition: We say that a configuration p_0, p_1, p_2, p_3 is *in* the block B if $p_i \in \langle Q_i \rangle$ for $i = 0, 1, 2, 3$. In other words, the point in \square representing our configuration is contained in $\langle B \rangle$. Sometimes we will say that this configuration is *associated to* the block.

Sudvidision of Blocks: There are 4 obvious subdivision operations we can perform on a block.

- The operation S_0 divides B into the two blocks (Q_{00}, Q_1, Q_2, Q_3) and (Q_{01}, Q_1, Q_2, Q_3) . Here (Q_{00}, Q_{01}) is the dyadic subdivision of Q_0 .
- the operation S_1 divides B into the 4 blocks (Q_0, Q_{1ab}, Q_2, Q_3) , where $(Q_{100}, Q_{101}, Q_{110}, Q_{111})$ is the dyadic subdivision of Q_1 .

The operations S_2 and S_3 are similar to S_1 .

The set of dyadic blocks has a natural poset structure to it, and basically our algorithm does a depth-first-search through this poset, eliminating solid blocks either due to symmetry considerations or due to energy considerations.

7.5 A Technical Result about Dyadic Boxes

The following result will be useful for the estimates in §8.4. We state it in more generality than we need, to show what hypotheses are required, but we note that good dyadic squares satisfy the hypotheses. We only care about the result for good dyadic squares and for good dyadic segments. In the case of good dyadic segments, the lemma is obviously true.

Lemma 7.5.1 *Let \widehat{Q} be a rectangle whose sides are parallel to the coordinate axes and do not cross the coordinate axes. Then the points of \widehat{Q} closest to $(0, 0, 1)$ and farthest from $(0, 0, 1)$ are both vertices.*

Proof: Put the metric on $\mathbf{R}^2 \cup \infty$ which makes stereographic projection an isometry. By symmetry, the metric balls about ∞ are the complements of disks centered at 0. The smallest disk centered at 0 and containing $\langle Q \rangle$ must have a vertex on its boundary. Likewise, the largest disk centered at 0 and disjoint from the interior of $\langle Q \rangle$ must have a vertex in its boundary. This latter result uses the fact that $\langle Q \rangle$ does not cross the coordinate axes. These statements are equivalent to the statement of the lemma. ♠

7.6 Very Near the TBP

Our calculations will depend on a pair (S, ϵ_0) , both powers of two. We always take $S = 2^{30}$ and $\epsilon = 2^{-18}$.

Define the *in-radius* of a cube to be half its side length. Let P_0 denote the configuration representing the totally normalized polar TBP and let B_0 denote the cube centered at P_0 and having in-radius ϵ_0 . Note that B_0 is not a dyadic block. This does not bother us. Here we give a sufficient condition for $B \subset B_0$, where B is some dyadic block.

Let $a = 1/\sqrt{3}$. For each choice of S we compute a value a^* such that $Sa^* \in \mathbf{Z}$ and $|a - a^*| < 1/S$. There are two such choices, namely

$$\frac{\text{floor}(Sa)}{S}, \quad \frac{\text{floor}(Sa + 1)}{S}. \quad (7.11)$$

In practice, our program sets a^* to be the first of these two numbers, but we want to state things in a symmetric way that works for either choice.

We define $B'_0 = Q'_0 \times Q'_1 \times Q'_2 \times Q'_3$ where

$$\begin{aligned} Q'_0 &= [1 - \epsilon_0, 1 + \epsilon_0], \\ Q'_1 &= [-\epsilon_0, \epsilon_0] \times [-S^{-1} - a^* - \epsilon_0, S^{-1} - a^* + \epsilon_0] \\ Q'_2 &= [-1 - \epsilon_0, -1 + \epsilon_0] \times [-\epsilon_0, \epsilon_0] \\ Q'_3 &= [-\epsilon_0, \epsilon_0] \times [-S^{-1} + a^* - \epsilon_0, S^{-1} + a^* + \epsilon_0] \end{aligned}$$

By construction, we have $B'_0 \subset B_0$. Given a block $B = Q_0 \times Q_1 \times Q_2 \times Q_3$, the condition $Q_i \subset Q'_i$ for all i implies that $B \subset B_0$. We will see in §10.6 that this is an exact integer calculation for us.

Chapter 8

Spherical Geometry Estimates

8.1 Overview

In this chapter we define the basic quantities that go into the Energy Theorem, Theorem 9.2.1. We will persistently use the hat and hull notation defined in §7.4. Thus:

- When Q is a dyadic square, $\langle \widehat{Q} \rangle$ is a convex quadrilateral in space whose vertices are 4 co-circular points on S^2 , and $\widehat{\langle Q \rangle}$ is a subset of S^2 bounded by 4 circular arcs.
- When Q is a dyadic segment, $\langle \widehat{Q} \rangle$ is a segment whose endpoints are on S^2 , and $\widehat{\langle Q \rangle}$ is an arc of a great circle.

Here is a summary of the quantities we will define in this chapter. The first three quantities are not rational functions of the inputs, but our estimates only use the squares of these quantities.

Hull Diameter: $d(Q)$ will be the diameter $\langle \widehat{Q} \rangle$.

Edge Length: $d_1(Q)$ will be the length of the longest edge of $\langle \widehat{Q} \rangle$.

Circular Measure: Let $D_Q \subset \mathbf{R}^2$ denote the disk containing Q in its boundary. $d_2(Q)$ is the diameter of \widehat{D}_Q .

Hull Separation Constant: $\delta(Q)$ will be a constant such that every point

in $\widehat{\langle Q \rangle}$ is within $\delta(Q)$ of a point of $\widehat{\langle Q \rangle}$. This quantity is a rational function of the coordinates of Q .

Dot Product Bounds: We will introduce a (finitely computable, rational) quantity $(Q \cdot Q')_{\max}$ which has the property that

$$V \cdot V' \leq (Q \cdot Q')_{\max}$$

for all $V \in \widehat{\langle Q \rangle} \cup \widehat{\langle Q \rangle}$ and $V' \in \widehat{\langle Q' \rangle} \cup \widehat{\langle Q' \rangle}$.

8.2 Some Results about Circles

Here we prove a few geometric facts about circles and stereographic projection.

Lemma 8.2.1 *Let Q be a good dyadic square or a dyadic segment. The circular arcs bounding $\widehat{\langle Q \rangle}$ lie in circles having diameter at least 1.*

Proof: Let Σ denote stereographic projection. In the dyadic segment case, $\widehat{\langle Q \rangle}$ lies in a great circle. In the good dyadic square case, each edge of $\langle Q \rangle$ lies on a line L which contains a point p at most $3/2$ from the origin. But $\Sigma^{-1}(p)$ is at least 1 unit from $(0, 0, 1)$. Hence $\Sigma^{-1}(L)$, which limits on $(0, 0, 1)$ and contains $\Sigma^{-1}(p)$, has diameter at least 1. The set $\Sigma^{-1}(L \cup \infty)$ is precisely the circle extending the relevant edge of $\widehat{\langle Q \rangle}$ ♠

Let $D \subset \mathbf{R}^2$ be a disk of radius $r \leq R$ centered at a point which is R units from the origin. Let \widehat{D} denote the corresponding disk on S^2 . We consider \widehat{D} as a subset of \mathbf{R}^3 and compute its diameter in with respect to the Euclidean metric on \mathbf{R}^3 .

Lemma 8.2.2

$$\text{diam}^2(\widehat{D}) = \frac{16r^2}{1 + 2r^2 + 2R^2 + (R^2 - r^2)^2}. \quad (8.1)$$

Proof: By symmetry it suffices to consider the case when the center of D is $(R, 0)$. The diameter is then achieved by the two points $V = \Sigma^{-1}(R - r, 0)$

and $W = \Sigma^{-1}(R + r, 0)$. The formula comes from computing $\|V - W\|^2$ and simplifying. ♠

We introduce the functions

$$\chi(D, d) = \frac{d^2}{4D} + \frac{d^4}{2D^3} \quad \chi^*(D, d) = \frac{1}{2}(D - \sqrt{D^2 - d^2}). \quad (8.2)$$

The second of these is a function closely related to the geometry of circles. This is the function we would use if we had an ideal computing machine. However, since we want our estimates to all be rational functions of the inputs, we will use the first function. We first prove an approximation lemma and then we get to the main point.

Lemma 8.2.3 *If $0 \leq d \leq D$ then $\chi^*(D, d) \leq \chi(D, d)$.*

Proof: If we replace (d, D) by (rd, rD) then both sides scale up by r . Thanks to this homogeneity, it suffices to prove the result when $D = 1$. We have $\chi(1, 1) = 3/4 > 1/2 = \chi^*(1, 1)$. So, it suffices to prove that the equation

$$\chi(1, d) - \chi^*(1, d) = \frac{d^2}{4} + \frac{d^4}{2} - \frac{1}{2}(1 - \sqrt{1 - d^2})$$

has no real solutions in $[0, 1]$ besides $d = 0$. Consider a solution to this equation. Rearranging the equation, we get $A = B$ where

$$A = -\frac{1}{2}\sqrt{1 - d^2}, \quad B = \frac{d^2}{4} + \frac{d^4}{2} - 1/2.$$

An exercise in calculus shows that the only roots of $A^2 - B^2$ in $[0, 1]$ are 0 and $\sqrt{1/2(\sqrt{8} - 1)} > .95$. On the other hand $A < 0$ and $B > 0$ on $[.95, 1]$. ♠

Now we get to the key result. This result holds in any dimension, but we will apply it once to the 2-sphere, and once to circles contained in suitable planes in \mathbf{R}^3 .

Lemma 8.2.4 *Let Γ be a round sphere of diameter D , contained in some Euclidean space. Let B be the ball bounded by Γ . Let Π be a hyperplane which intersects B but does not contain the center of B . Let $\gamma = \Pi \cap B$ and let γ^* be the smaller of the two spherical caps on Γ bounded by $\Pi \cap \Gamma$. Let $p^* \in \gamma^*$ be a point. Let $p \in \gamma$ be the point so the line $\overline{pp^*}$ contains the center of B . Then $\|p - p^*\| \leq \chi(D, d)$.*

Proof: The given distance is maximized when p^* is the center of γ^* and p is the center of γ . In this case it suffices by symmetry to consider the situation in \mathbf{R}^2 , where $\overline{pp^*}$ is the perpendicular bisector of γ . Setting $x = \|p - p^*\|$, we have

$$x(D - x) = (d/2)^2. \quad (8.3)$$

This equation comes from a well-known theorem from high school geometry concerning the lengths of crossing chords inside a circle. When we solve Equation 8.3 for x , we see that $x = \chi^*(D, d)$. The previous lemma finishes the proof. ♠

8.3 The Hull Approximation Lemma

Circular Measure: When Q is a dyadic square or segment, we define

$$d_2(Q) = \text{diam}(\widehat{D}_Q), \quad (8.4)$$

Where $D_Q \subset \mathbf{R}^2$ is such that $Q \subset D_Q$. So $d_2(Q)$ is the diameter of the small spherical cap which contains \widehat{Q} in its boundary. Note that $\widehat{\langle Q \rangle} \subset \widehat{D}_Q$ by construction. We call $d_2(Q)$ the *circular measure* of Q .

Hull Separation Constant: Recall that $d_1(Q)$ is the maximum side length of $\langle Q \rangle$. When Q is a dyadic segment, we define $\delta(Q) = \chi(2, d_2)$. When Q is a good dyadic square, We define

$$\delta(Q) = \max \left(\chi(1, d_1), \chi(2, d_2) \right). \quad (8.5)$$

This definition makes sense, because $d_1(Q) \leq 1$ and $d_2(Q) \leq \sqrt{2} < 2$. The point here is that Σ^{-1} is 2-Lipschitz and Q has side length at most $1/2$. We call $\delta(Q)$ the *Hull approximation constant* of Q .

Lemma 8.3.1 (Hull Approximation) *Let Q be a dyadic segment or a good dyadic square. Every point of the spherical patch $\widehat{\langle Q \rangle}$ is within $\delta(Q)$ of a point of the convex quadrilateral $\langle \widehat{Q} \rangle$.*

Proof: Suppose first that Q is a dyadic segment. $\widehat{\langle Q \rangle}$ is the short arc of a great circle sharing endpoints with $\langle \widehat{Q} \rangle$, a chord of length d_2 . By Lemma

8.2.4 each point of on the circular arc is within $\chi(2, d_2)$ of a point on the chord.

Now suppose that Q is a good dyadic square. Let O be the origin in \mathbf{R}^3 . Let $H \subset S^2$ denote the set of points such that the segment $Op^* \in H$ intersects $\langle \widehat{Q} \rangle$ in a point p . Here H is the intersection with the cone over $\langle \widehat{Q} \rangle$ with S^2 .

Case 1: Let $p^* \in \widehat{\langle Q \rangle} \cap H$. Let $p \in \langle Q \rangle$ be such that the segment Op^* contains p . Let B be the unit ball. Let Π be the plane containing \widehat{Q} . Note that $\Pi \cap S^2$ bounds the spherical \widehat{D}_Q which contains \widehat{Q} in its boundary. Let

$$\Gamma = S^2, \quad \gamma^* = \widehat{D}_Q, \quad \gamma = \Pi \cap B.$$

The diameter of Γ is $D = 2$. Lemma 8.2.4 now tells us that $\|p - p^*\| \leq \chi(2, d_2)$.

Case 2: Let $p^* \in \widehat{\langle Q \rangle} - H$. The sets $\widehat{\langle Q \rangle}$ and H are both bounded by 4 circular arcs which have the same vertices. H is bounded by arcs of great circles and $\widehat{\langle Q \rangle}$ is bounded by arcs of circles having diameter at least 1. The point p^* lies between an edge-arc α_1 of H and an edge-arc α_2 of $\widehat{\langle Q \rangle}$ which share both endpoints. Let γ be the line segment joining these endpoints. The diameter of γ is at most d_1 .

Call an arc of a circle *nice* if it is contained in a semicircle, and if the circle containing it has diameter at least 1. The arcs α_1 and α_2 are both nice. We can foliate the region between α_1 and α_2 by arcs of circles. These circles are all contained in the intersection of S^2 with planes which contain γ . Call this foliation \mathcal{F} . We get this foliation by rotating the planes around their common axis, which is the line through γ .

Say that an \mathcal{F} -circle is a circle containing an arc of \mathcal{F} . Let e be the edge of $\langle Q \rangle$ corresponding to γ . Call (e, Q) *normal* if γ is never the diameter of an \mathcal{F} -circle. If (e, Q) is normal, then the diameters of the \mathcal{F} -circles interpolate monotonically between the diameter of α_1 and the diameter of α_2 . Hence, all \mathcal{F} -circles have diameter at least 1. At the same time, if (e, Q) is normal, then all arcs of \mathcal{F} are contained in semicircles, by continuity. In short, if (e, Q) is normal, then all arcs of \mathcal{F} are nice. Assuming that (e, Q) is normal, let γ^* be the arc in \mathcal{F} which contains p^* . Let $p \in \Gamma$ be such that the line pp^* contains the center of the circle Γ containing γ^* . Since γ^* is nice, Lemma 8.2.4 says that $\|p - p^*\| \leq \chi(D, d_1) \leq \chi(1, d_1)$.

To finish the proof, we just have to show that (e, Q) is normal. We enlarge the set of possible pairs we consider, by allowing rectangles in $[-3/2, 3/2]^2$ having sides parallel to the coordinate axes and maximum side length $1/2$. The same arguments as above, Lemma 8.2.1 and the 2-Lipschitz nature of Σ^{-1} , show that α'_1 and α'_2 are still nice for any such pair (e', Q') .

If e' is the long side of a $1/2 \times 10^{-100}$ rectangle $\langle Q' \rangle$ contained in the 10^{-100} -neighborhood of the coordinate axes, then (e', Q') is normal: The arc α'_1 is very nearly the arc of a great circle and the angle between α'_1 and α'_2 is very small, so all arcs \mathcal{F}' are all nearly arcs of great circles. If some choice (e, Q) is not normal, then by continuity, there is a choice (e'', Q'') in which γ'' is the diameter of one of the two boundary arcs of \mathcal{F}'' . There is no other way to switch from normal to not normal. But this is absurd because the boundary arcs, α''_1 and α''_2 , are nice. ♠

8.4 Dot Product Estimates

Let Q be a dyadic segment or a good dyadic square. Let δ be the hull separation constant of Q . Let $\{q_i\}$ be the points of Q . We make all the same definitions for a second dyadic square Q' . We define

$$(Q \cdot Q')_{\max} = \max_{i,j} (\widehat{q}_i \cdot \widehat{q}'_j) + \delta + \delta' + \delta\delta'. \quad (8.6)$$

$$(Q \cdot \{\infty\})_{\max} = \max_i \widehat{q}_i \cdot (0, 0, 1) \quad (8.7)$$

Connectors: We say that a *connector* is a line segment connecting a point on \widehat{Q} to any of its closest points in $\widehat{Q'}$. We let $\Omega(Q)$ denote the set of connectors defined relative to Q . By the Hull Approximation Lemma, each $V \in \Omega(Q)$ has the form $W + \delta U$ where $W \in \widehat{Q}$ and $\|U\| \leq 1$.

Lemma 8.4.1 $V \cdot V' \leq (Q \cdot Q')_{\max}$ for all $(V, V') \in \Omega(Q) \times \Omega(Q')$

Proof: Suppose $V \in \widehat{Q}$ and $V' \in \widehat{Q'}$. Since the dot product is bilinear, the restriction of the dot product to the convex polyhedral set $\widehat{Q} \times \widehat{Q'}$ takes on its extrema at vertices. Hence $V \cdot V' \leq \max_{i,j} q_i \cdot q'_j$. In this case, we get the desired inequality whether or not $Q' = \{\infty\}$.

Suppose $Q' \neq \{\infty\}$ and V, V' are arbitrary. We use the decomposition mentioned above:

$$V = W + \delta U, \quad V' = W' + \delta' U, \quad W \in \langle \widehat{Q} \rangle, \quad W' \in \langle \widehat{Q'} \rangle. \quad (8.8)$$

But then, by the Cauchy-Schwarz inequality,

$$|(V \cdot V') - (W \cdot W')| = |V \cdot \delta' U' + V' \cdot \delta U + \delta U \cdot \delta' U'| \leq \delta + \delta' + \delta\delta'.$$

The lemma now follows immediately from this equation, the previous case applied to W, W' , and the triangle inequality.

Suppose that $Q' = \{\infty\}$. We already know the result when $V \in \langle \widehat{Q} \rangle$. When $V \in \widehat{\langle Q \rangle}$ we get the better bound above from Lemma 7.5.1 and from the fact that the dot product $V \cdot (0, 0, 1)$ varies monotonically with the distance from V to $(0, 0, 1)$ and *vice versa*. Now we know the result whenever V is the endpoint of a connector. By the linearity of the dot product, the result holds also when V is an interior point of a connector. ♠

8.5 Disordered Blocks

$B = (Q_0, Q_1, Q_2, Q_3)$ be a good block. As usual $Q_4 = \{\infty\}$. For each index j we define

$$(B, k)_{\min} = \min_{j \neq k} (Q_j \cdot Q_k)_{\min}, \quad (B, k)_{\max} = \min_{j \neq k} (Q_j \cdot Q_k)_{\max}. \quad (8.9)$$

We say that B is *disordered* if there is some $k \in \{0, 1, 2, 3\}$ such that

$$(B, 4)_{\max} < (B, k)_{\min} \quad (8.10)$$

Lemma 8.5.1 *If B is disordered, then no configuration in B is totally normalized.*

Proof: In this situation, every configuration $\{p_j\}$ in B has the following property.

$$\min_{i \neq 4} p_4 \cdot p_i < \min_{i \neq k} p_k \cdot p_j. \quad (8.11)$$

But square-distance is (linearly) decreasing as a function of the dot product. Hence

$$\delta(4) = \max_{i \neq 4} \|p_4 - p_i\|^2 > \max_{i \neq k} \|p_k - p_i\|^2 = \delta(k). \quad (8.12)$$

In short $\delta(4) > \delta(k)$ for every configuration associated to B . This is to say that none of these configurations is totally normalized. ♠

8.6 Irrelevant Blocks

Call a block *irrelevant* if no configuration in the interior of the block is totally normalized. Call a block *relevant* if it is not irrelevant. Every relevant configuration in the boundary of an irrelevant block is also in the boundary of a relevant block. So, to prove the Big and Small Theorems, we can ignore the irrelevant blocks.

Disordered blocks are irrelevant, but there are other irrelevant blocks. Here we give a criterion for a good block B to be irrelevant. Given a box Q_j , let \overline{Q}_{jk} and \underline{Q}_{jk} denote the maximum and minimum k th coordinate of a point in Q_j . If at least one of the following holds, The good block is irrelevant provided that at least one of the following holds.

1. $\min(|\underline{Q}_{k1}|, |\overline{Q}_{k1}|) \geq \overline{Q}_{01}$ for some $k = 1, 2, 3$.
2. $\min(|\underline{Q}_{k2}|, |\overline{Q}_{k2}|) \geq \overline{Q}_{01}$ for some $k = 1, 2, 3$.
3. $\underline{Q}_{12} \geq 0$, provided we have a monotone decreasing energy.
4. $\overline{Q}_{22} \leq 0$.
5. $\overline{Q}_{32} \leq \underline{Q}_{22}$.
6. $\overline{Q}_{22} \leq \underline{Q}_{12}$.
7. B is disordered.

Conditions 1 and 2 each imply that there is some index $k \in \{1, 2, 3\}$ such that all points in the interior of B_k are farther from the origin than all points in B_0 . Condition 3 implies that all points in the interior of B_1 lie above the x -axis. This violates our normalization when the energy function

is monotone decreasing. Condition 4 implies that all points in the interior of B_2 lie below the x -axis. Condition 5 implies that all points in the interior of B_3 lie below all points in the interior of B_2 . Condition 6 implies that all points in the interior of B_2 lie below all points in the interior of B_1 . We have already discussed Condition 7. Thus, if any of these conditions holds, the block is irrelevant.

Chapter 9

The Energy Theorem

9.1 Main Result

We think of the energy potential $G = G_k$ as being a function on $(\mathbf{R}^2 \times \infty)^2$, via the identification $p \leftrightarrow \hat{p}$. We take $k \geq 1$ to be an integer.

Let \mathcal{Q} denote the set of dyadic squares $[-2, 2]^2$ together with the dyadic segments in $[0, 4]$, together with $\{\infty\}$. When $Q = \{\infty\}$ the corresponding hull separation constant and the hull diameter are 0.

Now we are going to define a function $\epsilon : \mathcal{Q} \times \mathcal{Q} \rightarrow [0, \infty)$. First of all, for notational convenience we set $\epsilon(Q, Q) = 0$ for all Q . When $Q, Q' \in \mathcal{Q}$ are unequal, we define

$$\epsilon(Q, Q') = \frac{1}{2}k(k-1)T^{k-2}d^2 + 2kT^{k-1}\delta \quad (9.1)$$

Here

- d is the diameter of \hat{Q} .
- $\delta = \delta(Q)$ is the hull approximation constant for Q . See §8.3.
- $T = T(Q, Q') = 2 + 2(Q \cdot Q')_{\max}$. See §8.4.

This is a rational function in the coordinates of Q and Q' . The quantities d^2 and δ are essentially quadratic in the side-lengths of Q and Q' . Note that we have $\epsilon(\{\infty\}, Q') = 0$ but $\epsilon(Q, \{\infty\})$ is nonzero when $Q \neq \{\infty\}$.

Let $B = (Q_0, Q_1, Q_2, Q_3)$. For notational convenience we set $Q_4 = \{\infty\}$. We define

$$\mathbf{ERR}(B) = \sum_{i=0}^3 \sum_{j=0}^4 \epsilon(Q_i, Q_j). \quad (9.2)$$

Theorem 9.1.1

$$\min_{v \in \langle B \rangle} \mathcal{E}_k(v) > \min_{v \in B} \mathcal{E}_k(v) - \mathbf{ERR}(B), \quad \max_{v \in \langle B \rangle} \mathcal{E}_k(v) < \max_{v \in B} \mathcal{E}_k(v) + \mathbf{ERR}(B).$$

Remark: A careful examination of our proof will reveal that, for the case of max, one can get away with just using the second term in the definition of ϵ in Equation 9.1. However, we are so rarely interested in this case that we will let things stand as they are.

9.2 Generalizations and Consequences

Theorem 9.1.1 suffices to deal with G_3, G_4, G_6 , but we need a more general result to deal with G_5^b , G_{10}^\sharp and $G_{10}^{\sharp\sharp}$. Suppose we have some energy of the form

$$F = \sum_{k=1}^N a_k G_k \quad (9.3)$$

where a_1, \dots, a_N is some sequence of numbers, not necessarily positive.

Suppressing the dependence on F , we define

$$\epsilon(Q_i, Q_j) = \sum |a_k| \epsilon_k(Q_i, Q_j), \quad (9.4)$$

where $\epsilon_k(Q_i, Q_j)$ is the above quantity computed with respect to G_k . We then define **ERR** exactly as in Equation 9.2. Here is an immediate consequence of Theorem 9.1.1. (The other statement of Theorem 9.1.1 holds as well, but we don't care about it.)

Theorem 9.2.1 (Energy)

$$\min_{v \in \langle B \rangle} \mathcal{E}(v) > \min_{v \in B} \mathcal{E}(v) - \mathbf{ERR}(B).$$

Proof: Choose some $v \in B$ and $z \in \langle B \rangle$. By the triangle inequality and Theorem 9.1.1,

$$|\mathcal{E}(v) - \mathcal{E}(z)| \leq \sum |a_k| |\mathcal{E}_k(v) - \mathcal{E}_k(z)| \leq \sum |a_k| \mathbf{ERR}_k = \mathbf{ERR}. \quad (9.5)$$

This holds for all pairs v, z . The conclusion of the Energy Theorem follows immediately. ♠

Corollary 9.2.2 *Suppose that B is a block such that*

$$\min_{v \in B} \mathcal{E}(v) - \mathbf{ERR}(B) > \mathcal{E}(\text{TBP}). \quad (9.6)$$

Then all configurations in B have higher energy than the TBP.

We can write

$$\mathbf{ERR}(B) = \sum_{i=0}^3 \mathbf{ERR}_i(B), \quad \mathbf{ERR}_i(B) = \sum_{j=0}^4 \epsilon(Q_i, Q_j). \quad (9.7)$$

We define the *subdivision recommendation* to be the index $i \in \{0, 1, 2, 3\}$ for which $\mathbf{ERR}_i(B)$ is maximal. In the extremely unlikely event that two of these terms coincide, we pick the smaller of the two indices to break the tie. The subdivision recommendation feeds into the algorithm described in §10.

9.3 A Polynomial Inequality

The rest of the chapter is devoted to proving Theorem 9.1.1. One building block of Theorem 9.1.1 is the case $M = 4$ of the following inequality.

Lemma 9.3.1 *Let $M \geq 2$ and $k = 1, 2, 3, \dots$. Suppose*

- $0 \leq x_1 \leq \dots \leq x_M$.
- $\sum_{i=1}^M \lambda_i = 1$ and $\lambda_i \geq 0$ for all i .

Then

$$0 \leq \sum_{i=1}^M \lambda_i x_i^k - \left(\sum_{i=1}^M \lambda_i x_i \right)^k \leq \frac{1}{8} k(k-1) x_M^{k-2} (x_M - x_1)^2. \quad (9.8)$$

The lower bound is a trivial consequence of convexity, and both bounds are trivial when $k = 1$. So, we take $k = 2, 3, 4, \dots$ and prove the upper bound. I discovered Lemma 9.3.1 experimentally.

Lemma 9.3.2 *The case $M = 2$ of Lemma 9.3.1 implies the rest.*

Proof: Suppose that $M \geq 3$. We have one degree of freedom when we keep $\sum \lambda_i x_i$ constant and try to vary $\{\lambda_j\}$ so as to maximize the left hand side of the inequality. The right hand side does not change when we do this, and the left hand side varies linearly. Hence, the left hand size is maximized when $\lambda_i = 0$ for some i . But then any counterexample to the lemma for $M \geq 3$ gives rise to a counter example for $M - 1$. ♠

In the case $M = 2$, we set $a = \lambda_1$. Both sides of the inequality in Lemma 9.3.1 are homogeneous of degree k , so it suffices to consider the case when $x_2 = 1$. We set $x = x_1$. The inequality of is then $f(x) \leq g(x)$, where

$$f(x) = (ax^k + 1 - a) - (ax + 1 - a)^k; \quad g(x) = \frac{1}{8}k(k-1)(1-x)^2. \quad (9.9)$$

This is supposed to hold for all $a, x \in [0, 1]$.

The following argument is due to C. McMullen, who figured it out after I told him about the inequality.

Lemma 9.3.3 *Equation 9.9 holds for all $a, x \in [0, 1]$ and all $k \geq 2$.*

Proof: Equation 9.9. We think of f as a function of x , with a held fixed. Since $f(1) = g(1) = 1$, it suffices to prove that $f'(x) \geq g'(x)$ on $[0, 1]$. Define

$$\phi(x) = akx^{k-1}, \quad b = (1-a)(1-x). \quad (9.10)$$

We have

$$-f'(x) = \phi(x+b) - \phi(x). \quad (9.11)$$

Both x and $x+b$ lie in $[0, 1]$. So, by the mean value theorem there is some $y \in [0, 1]$ so that

$$\frac{\phi(x+b) - \phi(x)}{b} = \phi'(y) = ak(k-1)y^{k-2}. \quad (9.12)$$

Hence

$$-f'(x) = b\phi'(y) = a(1-a)k(k-1)(1-x)y^{k-2} \quad (9.13)$$

But $a(1-a) \in [0, 1/4]$ and $y^{k-2} \in [0, 1]$. Hence

$$-f'(x) \leq \frac{1}{4}k(k-1)(1-x) = -g'(x). \quad (9.14)$$

Hence $f'(x) \geq g'(x)$ for all $x \in [0, 1]$. ♠

Remark: Lemma 9.3.1 has the following motivation. The idea behind the Energy Theorem is that we want to measure the deviation of the energy function from being linear, and for this we would like a quadratic estimate. Since our energy G_k involves high powers, we want to estimate these high powers by quadratic terms.

9.4 The Local Energy Lemma

Let $Q = \{q_1, q_2, q_3, q_4\}$ be the vertex set of $Q \in \mathcal{Q}$. We allow for the degenerate case that Q is a line segment or $\{\infty\}$. In this case we just list the vertices multiple times, for notational convenience.

Note that every point in the convex quadrilateral $\langle \widehat{Q} \rangle$ is a convex average of the vertices. For each $z \in \langle Q \rangle$, there is a some point $z^* \in \langle \widehat{Q} \rangle$ which is as close as possible to $\widehat{z} \in \langle \widehat{Q} \rangle$. There are constants $\lambda_i(z)$ such that

$$z^* = \sum_{i=1}^4 \lambda_i(z) \widehat{q}_i, \quad \sum_{i=1}^4 \lambda_i(z) = 1. \quad (9.15)$$

We think of the 4 functions $\{\lambda_i\}$ as a partition of unity on $\langle Q \rangle$. The choices above might not be unique, but we make such choices once and for all for each Q . We call the assignment $Q \rightarrow \{\lambda_i\}$ the *stereographic weighting system*.

Lemma 9.4.1 (Local Energy) *Let ϵ be the function defined in the Theorem 9.1.1. Let Q, Q' be distinct members of \mathcal{Q} . Given any $z \in Q$ and $z' \in Q'$,*

$$\left| G(z, z') - \sum_{i=1}^4 \lambda_i(z) G(q_i, z') \right| \leq \epsilon(Q, Q'). \quad (9.16)$$

Proof: For notational convenience, we set $w = z'$. Let

$$X = (2 + 2z^* \cdot \widehat{w})^k. \quad (9.17)$$

The Local Energy Lemma follows from the triangle inequality and the following two inequalities

$$\left| \sum_{i=1}^4 \lambda_i G(q_i, w) - X \right| \leq \frac{1}{2} k(k-1) T^{k-2} d^2 \quad (9.18)$$

$$|X - G(z, w)| \leq 2k T^{k-1} \delta. \quad (9.19)$$

We will establish these inequalities in turn.

Let q_1, q_2, q_3, q_4 be the vertices of Q . Let $\lambda_i = \lambda_i(z)$. We set

$$x_i = 4 - \|\widehat{q}_i - \widehat{w}\|^2 = 2 + 2\widehat{q}_i \cdot \widehat{w}, \quad i = 1, 2, 3, 4. \quad (9.20)$$

Note that $x_i \geq 0$ for all i . We order so that $x_1 \leq x_2 \leq x_3 \leq x_4$. We have

$$\sum_{i=1}^4 \lambda_i(z) G(q_i, w) = \sum_{i=1}^4 \lambda_i x_i^k, \quad (9.21)$$

$$X = (2 + 2\widehat{z}^* \cdot \widehat{w})^k = \left(\sum_{i=1}^4 \lambda_i (2 + \widehat{q}_i \cdot \widehat{w}) \right)^k = \left(\sum_{i=1}^4 \lambda_i x_i \right)^k. \quad (9.22)$$

By Equation 9.21, Equation 9.22, and the case $M = 4$ of Lemma 9.3.1,

$$\left| \sum_{i=1}^4 \lambda_i G(q_i, w) - X \right| = \left| \sum_{i=1}^4 \lambda_i x_i^k - \left(\sum_{i=1}^4 \lambda_i x_i \right)^k \right| \leq \frac{1}{8} k(k-1) x_4^{k-2} (x_4 - x_1)^2. \quad (9.23)$$

By Lemma 8.4.1, we have

$$x_4 = 2 + 2(\widehat{q}_4 \cdot \widehat{w}) \leq 2 + 2(Q \cdot Q')_{\max} = T. \quad (9.24)$$

Since d is the diameter of $\langle \widehat{Q} \rangle$ and \widehat{w} is a unit vector,

$$x_4 - x_1 = 2\widehat{w} \cdot (\widehat{q}_4 - \widehat{q}_1) \leq 2\|\widehat{w}\| \|\widehat{q}_4 - \widehat{q}_1\| = 2\|\widehat{q}_4 - \widehat{q}_1\| \leq 2d. \quad (9.25)$$

Plugging Equations 9.24 and 9.25 into Equation 9.23, we get Equation 9.18.

Now we establish Equation 9.19. Let γ denote the unit speed line segment connecting \widehat{z} to z^* . Note that the length L of γ is at most δ , by the Hull Approximation Lemma. Define

$$f(t) = \left(2 + 2\widehat{w} \cdot \gamma(t)\right)^k. \quad (9.26)$$

We have $f(0) = X$. Since \widehat{w} and $\gamma(1) = \widehat{z}$ are unit vectors, $f(L) = G(z, w)$. Hence

$$X - G(z, w) = f(0) - f(L), \quad L \leq \delta. \quad (9.27)$$

By the Chain Rule,

$$f'(t) = (2\widehat{w} \cdot \gamma'(t)) \times k \left(2 + 2\widehat{w} \cdot \gamma(t)\right)^{k-1}. \quad (9.28)$$

Note that $|2\widehat{w} \cdot \gamma'(t)| \leq 2$ because both of these vectors are unit vectors. γ parametrizes one of the connectors from Lemma 8.4.1, so

$$|f'(t)| \leq 2k \left(2 + 2\widehat{w} \cdot \gamma(t)\right)^{k-1} \leq 2k \left(2 + 2(Q \cdot Q')_{\max}\right)^{k-1} = 2kT^{k-1}. \quad (9.29)$$

Equation 9.19 now follows from Equation 9.27, Equation 9.29, and integration. ♠

9.5 From Local to Global

Let ϵ be the function from the Energy Theorem. Let $B = (Q_0, \dots, Q_N)$ be a list of $N + 1$ elements of \mathcal{Q} . We care about the case $N = 4$ and $Q_4 = \{\infty\}$, but the added generality makes things clearer. Let $q_{i,1}, q_{i,2}, q_{i,3}, q_{i,4}$ be the vertices of Q_i . The vertices of $\langle B \rangle$ are indexed by a multi-index

$$I = (i_0, \dots, i_n) \in \{1, 2, 3, 4\}^{N+1}.$$

Given such a multi-index, which amounts to a choice of vertex of $\langle B \rangle$, we define the energy of the corresponding vertex configuration:

$$\mathcal{E}(I) = \mathcal{E}(q_{0,i_0}, \dots, q_{N,i_N}) \quad (9.30)$$

We will prove the following sharper result.

Theorem 9.5.1 *Let $z_0, \dots, z_N \in \langle B \rangle$. Then*

$$\left| \mathcal{E}(z_0, \dots, z_N) - \sum_I \lambda_{i_0}(z_0) \dots \lambda_{i_N}(z_N) \mathcal{E}(I) \right| \leq \sum_{i=0}^N \sum_{j=0}^N \epsilon(Q_i, Q_j). \quad (9.31)$$

The sum is taken over all multi-indices.

Lemma 9.5.2 *Theorem 9.5.1 implies Theorem 9.1.1.*

Proof: Notice that

$$\sum_I \lambda_{i_0}(z_0) \dots \lambda_{i_N}(z_N) = \prod_{j=0}^N \left(\sum_{a=1}^4 \lambda_a(z_j) \right) = 1. \quad (9.32)$$

Therefore

$$\min_{v \in B} \mathcal{E}(v) \leq \sum_I \lambda_{i_0}(z_0) \dots \lambda_{i_N}(z_N) \mathcal{E}(I) \leq \max_{v \in B} \mathcal{E}(v), \quad (9.33)$$

because the sum in the middle is the convex average of vertex energies.

We will deal with the min case of Theorem 9.1.1. The max case has the same treatment. Choose some $(z_1, \dots, z_N) \in B$ which minimizes \mathcal{E} . We have

$$\begin{aligned} 0 &\leq \min_{v \in B} \mathcal{E}(v) - \min_{v \in \langle B \rangle} \mathcal{E}(v) = \min_{v \in B} \mathcal{E}(v) - \mathcal{E}(z_0, \dots, z_N) \leq \\ &\sum_I \lambda_{i_0}(z_0) \dots \lambda_{i_N}(z_N) \mathcal{E}(I) - \mathcal{E}(z_0, \dots, z_N) \leq \sum_{i=0}^N \sum_{j=0}^N \epsilon(Q_i, Q_j). \end{aligned} \quad (9.34)$$

The last expression is **ERR** when $N = 4$ and $Q_4 = \infty$. ♠

We now prove Theorem 9.5.1.

A Warmup Case

Consider the case when $N = 1$. Setting $\epsilon_{ij} = \epsilon(Q_i, Q_j)$, the Local Energy Lemma gives us

$$G(z_0, z_1) \geq \sum_{\alpha=1}^4 \lambda_{\alpha}(z_0) G(q_{0\alpha}, z_1) - \epsilon_{01}. \quad (9.35)$$

$$G(q_{0\alpha}, z_1) \geq \sum_{\beta=1}^4 \lambda_{\beta}(z_1) G(q_{1\beta}(z_1), q_{0\alpha}) - \epsilon_{10}. \quad (9.36)$$

Plugging the second equation into the first and using $\sum \lambda_{\alpha}(z_0) = 1$, we have

$$G(z_0, z_1) \geq \left(\sum_{\alpha, \beta} \lambda_{\alpha}(z_0) \lambda_{\beta}(z_1) G(q_{0\alpha}, q_{1\beta}) \right) - (\epsilon_{01} + \epsilon_{10}). \quad (9.37)$$

Similarly,

$$G(z_0, z_1) \leq \left(\sum_{\alpha, \beta} \lambda_{\alpha}(z_0) \lambda_{\beta}(z_1) G(q_{0\alpha}, q_{1\beta}) \right) + (\epsilon_{01} + \epsilon_{10}). \quad (9.38)$$

Equations 9.37 and 9.38 are equivalent to Equation 9.31 when $N = 1$.

The General Case

Now assume that $N \geq 2$. We rewrite Equation 9.37 as follows:

$$G(z_0, z_1) \geq \sum_A \lambda_{A_0}(z_0) \lambda_{A_1}(z_1) G(q_{0A_0}, q_{1A_1}) - (\epsilon_{01} + \epsilon_{10}). \quad (9.39)$$

The sum is taken over multi-indices A of length 2.

We also observe that

$$\sum_{I'} \lambda_{i_2}(z_2) \dots \lambda_{i_N}(z_N) = 1. \quad (9.40)$$

The sum is taken over all multi-indices $I' = (i_2, \dots, i_N)$. Therefore, if A is held fixed, we have

$$\lambda_{A_0}(z_0) \lambda_{A_1}(z_1) = \sum_I \lambda_{I_0}(z_0) \dots \lambda_{I_N}(z_N). \quad (9.41)$$

The sum is taken over all multi-indices of length $N + 1$ which have $I_0 = A_0$ and $I_1 = A_1$. Combining these equations, we have

$$G(z_0, z_1) \geq \sum_I \lambda_{I_0}(z_0) \dots \lambda_{I_N}(z_N) G(q_{0I_0}, q_{1I_1}) - (\epsilon_{01} + \epsilon_{10}). \quad (9.42)$$

The same argument works for other pairs of indices, giving

$$G(z_i, z_j) \geq \sum_I \lambda_{I_0}(z_0) \dots \lambda_{I_N}(z_N) G(q_{iI_i}, q_{jI_j}) - (\epsilon_{ij} + \epsilon_{ji}). \quad (9.43)$$

Now we interchange the order of summation and observe that

$$\begin{aligned}
\sum_{i < j} \left(\sum_I \lambda_{I_0}(z_0) \dots \lambda_{I_N}(z_N) G(q_{iI_i}, q_{jI_j}) \right) &= \\
\sum_I \sum_{i < j} \lambda_{I_0}(z_0) \dots \lambda_{I_N}(z_N) G(q_{iI_i}, q_{jI_j}) &= \\
\sum_I \lambda_{I_0}(z_0) \dots \lambda_{I_N}(z_N) \left(\sum_{i < j} G(q_{iI_i}, q_{jI_j}) \right) &= \\
\sum_I \lambda_{I_0}(z_0) \dots \lambda_{I_N}(z_N) \mathcal{E}(I). & \tag{9.44}
\end{aligned}$$

When we sum Equation 9.43 over all $i < j$, we get

$$\mathcal{E}(z_0, \dots, z_N) \geq \sum_I \lambda_{i_0}(z_0) \dots \lambda_{i_N}(z_N) \mathcal{E}(I) - \sum_{i=0}^N \sum_{j=0}^N \epsilon(Q_i, Q_j). \tag{9.45}$$

Similary,

$$\mathcal{E}(z_0, \dots, z_N) \leq \sum_I \lambda_{i_0}(z_0) \dots \lambda_{i_N}(z_N) \mathcal{E}(I) + \sum_{i=0}^N \sum_{j=0}^N \epsilon(Q_i, Q_j). \tag{9.46}$$

These two equations together are equivalent to Theorem 9.5.1. This completes the proof.

Chapter 10

The Main Algorithm

10.1 Grading a Block

In this section we describe what we mean by *grading* a block. The algorithm depends on the constants (S, ϵ_0) from §7.6. We take $S = 2^{30}$ and $\epsilon_0 = 2^{-18}$.

Let B_0 denote the cube of in-radius ϵ_0 about configuration of \square representing the normalized TBP. We fix some energy G_k . We perform the following tests on a block $B = (Q_0, Q_1, Q_2, Q_3)$, in the order listed.

1. If some component square Q_i of B has side length more than $1/2$ we fail B and recommend that B be subdivided along the first such index. This step guarantees that we only pass good blocks.
2. If B satisfies the criteria in §8.6, and hence is irrelevant, we pass B .
3. If we compute that $Q_i \not\subset [-3/2, 3/2]^2$ for some $i = 1, 2, 3$, we pass B . Given the previous step, Q_i is disjoint from $(-3/2, 3/2)^2$.
4. If the calculations in §7.6 show that $B \subset B_0$, we pass B .
5. If the calculations in §9.1 show that B satisfies Corollary 9.2.2, we pass B . Otherwise, we fail B and pass along the recommended subdivision.

10.2 The Divide and Conquer Algorithm

Now we describe the divide-and-conquer algorithm.

1. Begin with a list LIST of blocks in \square . Initially LIST consists of a single element, namely \square .
2. Let B be the last member of LIST. We delete B from LIST and then we grade B .
3. Suppose B passes. If LIST is empty, we halt and declare success. Otherwise, we return to Step 2.
4. Suppose B fails. In this case, we subdivide B along the subdivision recommendation and we append to LIST the subdivision of B . Then we return to Step 2.

If the algorithm halts with success, it implies that every relevant block B either lies in B_0 or does not contain a minimizer.

10.3 Results of the Calculations

For $G_3, G_4, G_5^b, G_6, G_{10}^{\#\#}$, I ran the programs (most recently) in late September 2016, on my 2014 IMAC.

- For G_3 the program finished in about 1 hour.
- For G_4 the program finished in about 1 hour and 9 minutes.
- For G_5^b the program finished in about 5 hours and 40 minutes.
- For G_6 the program finished in about 3 hours and 6 minutes.
- For $G_{10}^{\#\#}$ the program finished in about 25 hours and 32 minutes.

These programs are about 5 times as fast without the interval arithmetic.

In each case, the program produces a partition of \square into N_k smaller blocks, each of which is either irrelevant, contains no minimizer, or lies in B_0 . Here

$$(N_3, N_4, N_5^b, N_6, N_{10}^{\#\#}) = (4800136, 5302730, 13247122, 13212929, 41556654).$$

These calculations rigorously establish the following result.

Lemma 10.3.1 *Let $B_0 \subset \square$ denote the cube of in-radius 2^{-18} about the equatorial version of the TBP. If $P \in \square$ is a minimizer with respect to any of $G_3, G_4, G_5^\flat, G_6, G_{10}^\sharp$ then P is represented by a configuration in B_0 .*

Now we turn to the calculation with G_{10}^\sharp . For reference, here are the inequalities defining **SMALL**.

1. $\|p_0\| \geq \|p_k\|$ for $k = 1, 2, 3$.
2. $512p_0 \in [433, 498] \times [0, 0]$.
3. $512p_1 \in [-16, 16] \times [-464, -349]$.
4. $512p_2 \in [-498, -400] \times [0, 24]$.
5. $512p_3 \in [-16, 16] \times [349, 364]$.

Conditions 2-5 are obviously linear inequalities involving dyadic rationals. To test when a block lies in **SMALL** we are just testing linear inequalities for 128 vertices. Given the dyadic nature of the coefficients, this turns out to be an exact integer calculation for us.

We run the algorithm above for G_{10}^\sharp except that we replace B_0 in Step 4 of the grading with the set $B_0 \cup \mathbf{SMALL}$. That is, we first check if $B \subset B_0$, as before, and then we check if $B \subset \mathbf{SMALL}$. The program finished in about 35 hours and 42 minutes, and produced a partition of size 56256273. This rigorously establishes the following result.

Lemma 10.3.2 *Let $B_0 \subset \square$ denote the cube of in-radius 2^{-18} about the equatorial version of the TBP. If $P \in \square$ is a minimizer with respect to G_{10}^\sharp then either P is represented by a configuration in **SMALL** or by a configuration in B_0 .*

The rest of the chapter discusses the implementation of the algorithm.

10.4 Subdivision Recommendation

We first discuss a subtle point about the way the subdivision algorithm runs. The Energy Theorem is designed to run with the subdivision algorithm. As discussed at the end of §9.2 and mentioned again above in the description of the algorithm, the Energy Theorem comes with a recommendation for how

to subdivide the block in the way that is most likely to reduce the error term. Call this the *sophisticated approach*. The sophisticated approach might not always subdivide the longest side of the block.

As an alternative, we could perform the subdivisions just we se did for the positive dominance algorithm, always dividing so as to cut the longest side in half. Call this the *naive method*.

The sophisticated method is more efficient. As an experiment, I ran the program for G_3 , using floating point calculations, using the two approaches. The sophisticated approach completed in 16 minutes and produced a partition of size 4799998. The naive approach completed in 32 minutes and produced a partition of size 11372440.

10.5 Interval Arithmetic

As we mentioned in the introduction, everything we need to compute is a rational function of the coordinates of the block vertices. Thus, the calculations only involve the operations plus, minus, times, divide. In theory, they could be done with integer arithmetic. This seems to slow, so we do the calculations using interval arithmetic. Our implementation is like the in [S1] but here we don't need to worry about the square-root function.

Java represents real numbers by **doubles**, essentially according to the scheme discussed in [I, §3.2.2]. A double is a 64 bit string where 11 of the bits control the exponent, 52 of the bits control the binary expansion, and one bit controls the sign. The non-negative doubles have a lexicographic ordering, and this ordering coincides with the usual ordering of the real numbers they represent. The lexicographic ordering for the non-positive doubles is the reverse of the usual ordering of the real numbers they represent. To *increment* x_+ of a positive double x is the very next double in the ordering. This amounts to treating the last 63 bits of the string as an integer (written in binary) and adding 1 to it. With this interpretation, we have $x_+ = x + 1$. We also have the decrement $x_- = x - 1$. Similar operations are defined on the non-positive doubles. These operations are not defined on the largest and smallest doubles, but our program never encounters (or comes anywhere near) these.

Let \mathbf{D} be the set of all doubles. Let

$$\mathbf{R}_0 = \{x \in \mathbf{R} \mid |x| \leq 2^{50}\} \quad (10.1)$$

Our choice of 2^{50} is an arbitrary but convenient cutoff. Let \mathbf{D}_0 denote the set of doubles representing reals in \mathbf{R}_0 .

According to [I, 3.2.2, 4.1, 5.6], there is a map $\mathbf{R}_0 \rightarrow \mathbf{D}_0$ which maps each $x \in \mathbf{R}_0$ to some $[x] \in \mathbf{D}_0$ which is closest to x . In case there are several equally close choices, the computer chooses one according to the method in [I, §4.1]. This “nearest point projection” exists on a subset of \mathbf{R} that is much larger than \mathbf{R}_0 , but we only need to consider \mathbf{R}_0 . We also have the inclusion $r : \mathbf{D}_0 \rightarrow \mathbf{R}_0$, which maps a double to the real that it represents.

The basic operations plus, minus, times, divide act on \mathbf{R}_0 in the usual way, and operations with the same name act on \mathbf{D}_0 . Regarding these operations, [I, §5] states that *each of the operations shall be performed as if it first produced an intermediate result correct to infinite precision and with unbounded range, and then coerced this intermediate result to fit into the destination’s format*. Thus, for doubles $x, y \in \mathbf{D}_0$ such that $x * y \in \mathbf{D}_0$ as well, we have

$$x * y = [r(x) * r(y)]; \quad * \in \{+, -, \times, \div\}. \quad (10.2)$$

The operations on the left hand side represent operations on doubles and the operations on the right hand side represent operations on reals.

It might happen that $x, y \in \mathbf{D}_0$ but $x * y$ is not. To avoid this problem, we make the following checks before performing any arithmetic operation.

- For addition and subtraction, $\max(|x|, |y|) \leq 2^{40}$.
- For multiplication, $\max(|x|, |y|) < 2^{40}$ and $\min(|x|, |y|) < 2^{10}$.
- For division, $|x| \leq 2^{40}$ and $|y| \leq 2^{10}$ and $|y| \geq 2^{-10}$.

We set the calculation to abort if any of these conditions fails.

For us, an *interval* is a pair $I = (x, y)$ of doubles with $x \leq y$ and $x, y \in \mathbf{D}_0$. Say that I *bounds* $z \in \mathbf{R}_0$ if $x \leq [z] \leq y$. This is true if and only if $x \leq z \leq y$. Define

$$[x, y]_o = [x_-, y_+]. \quad (10.3)$$

This operation is well defined for doubles in \mathbf{D}_0 . We are essentially *rounding out* the endpoints of the interval. Let I_0 and I_1 denote the left and right endpoints of I . Letting I and J be intervals, we define

$$I * J = (\min_{ij} I_i * I_j, \max_{ij} I_i * I_j)_o. \quad (10.4)$$

That is, we perform the operations on all the endpoints, order the results, and then round outward. Given Equation 10.2, we the interval $I * J$ bounded $x * y$ provided that I bounds x and J bounds y . This operation is very similar to what we did in §4.1, except that here we are rounding out to take care of the round-off error. This rounding out process potentially destroys the semi-ring structure discussed in §4.1 but, as we remarked there, this doesn't bother us.

We also define an interval version of a vector in \mathbf{R}^3 . Such a vector consists of 3 intervals. The only operations we perform on such objects are addition, subtraction, scaling, and taking the dot product. These operations are all built out of the arithmetic operations.

All of our calculations come down to proving inequalities of the form $x < y$. We imagine that x and y are the outputs of some finite sequence of arithmetic operations and along the way we have intervals I_x and I_y which respectively bound x and y . If we know that the right endpoint of I_x is less than the left endpoint of I_y , then this constitutes a proof that $x < y$. The point is that the whole interval I_x lies to the left of I_y on the number line.

10.6 Integer Calculations

Now let me discuss the implementation of the divide and conquer algorithm. We manipulate blocks and dyadic squares using **longs**. These are 64 bit integers. Given a dyadic square Q with center (x, y) and side length 2^{-k} , we store the triple

$$(Sx, Sy, k). \quad (10.5)$$

Here $S = 2^{25}$ when we do the calculations for G_3, G_4, G_5, G_6 and $S = 2^{30}$ when we do the calculation for \hat{G}_{10} . The reader can modify the program so that it uses any power of 2 up to 2^{40} . Similarly, we store a dyadic segment with center x and side length 2^{-k} as (Sx, k) .

The subdivision is then obtained by manipulating these triples. For instance, the top right square in the subdivision of (Sx, Sy, k) is

$$(Sx - 2^{-k+1}S, Sy - 2^{-k+1}S, k + 1).$$

The scale 2^N allows for N such subdivisions before we lose the property that the squares are represented by integer triples. The biggest dyadic square is stored as $(0, 0, -2)$, and each subdivision increases the value of k by 1. We

terminate the algorithm if we ever arrive at a dyadic square whose center is not an even pair of integers. We never reach this eventuality when we run the program on the functions from the Main Theorem, but it does occur if we try functions like G_7 .

Steps 1,3, and 4 of the block grading just involve integer calculations. For instance, the point of scaling our square centers by S is that the inequalities which go into the calculations in §7.6 are all integer inequalities. We are simply clearing denominators.

Most of Step 2 just requires exact integer calculations. The only part of Step 2 that requires interval arithmetic calculations is the check that a block could be disordered, as in §8.5. In this case, we convert each long L representing a coordinate of our block into the dyadic interval

$$[L/S, L/S], \quad L = 2^{30}. \quad (10.6)$$

We then perform the calculations using interval arithmetic. The first step in the calculation is to apply inverse stereographic projection, as in Equation 7.1. The remaining steps are as in §8.5.

Step 5 is the main step that requires interval arithmetic. We do exactly the same procedure, and then perform the calculations which go into the Energy Theorem using interval arithmetic.

10.7 A Gilded Approach

In the interest of speed, we take what might be called a gilded approach to interval arithmetic. We run the calculations using integer arithmetic and floating point arithmetic until the point where we have determined that a block passes for a reason that requires a floating point calculation. Then we re-do the calculation using the interval arithmetic. This approach runs a bit faster, and only integer arithmetic and interval arithmetic are used to actually pass blocks.

10.8 A Potential Speed Up

One thing that probably slows our algorithm down needlessly is the computation of the constants d^2 and δ from Equation 9.1. These functions depend on the associated dyadic square or segment in a fairly predictable way, and

it ought to be possible to replace the calculations of d^2 and δ with an *a priori* estimate. Such estimates might also make it more feasible to run exact integer versions of the programs, because the sizes of the rational numbers in d^2 and δ get very large. I haven't tried to do this yet, but perhaps a later version of the program will do this.

10.9 Debugging

One serious concern about any computer-assisted proof is that some of the main steps of the proof reside in computer programs which are not printed, so to speak, along with the paper. It is difficult for one to directly inspect the code without a serious time investment, and indeed the interested reader would do much better simply to reproduce the code and see that it yields the same results.

The worst thing that could happen is if the code had a serious bug which caused it to suggest results which are not actually true. Let me explain the extent to which I have debugged the code. Each of the java programs has a debugging mode, in which the user can test that various aspects of the program are running correctly. While the debugger does not check every method, it does check that the main ones behave exactly as expected.

Here are some debugging features of the program.

- You can check on random inputs that the interval arithmetic operations are working properly.
- You can check on random inputs that the vector operations - dot product, addition, etc. - are working properly.
- You can check for random dyadic squares that the floating point and interval arithmetic measurements match in the appropriate sense.
- You can select a block of your choice and compare the estimate from the Energy Theorem with the minimum energy taken over a million random configurations in the block.
- You can open up an auxiliary window and see the grading step of the algorithm performed and displayed for a block of your choosing.

Chapter 11

Local Analysis of the Hessian

The purpose of this chapter is to prove, in all relevant cases that the Hessian of the energy function \mathcal{E} is positive definite throughout the neighborhood B_0 mentioned in Lemmas 10.3.1 and 10.3.2. This result finishes the proof of the Big Theorem and the Small Theorem.

11.1 Outline

We begin with a well-known lemma about the Hessian derivative of a function. The Hessian is the matrix of second partial derivatives.

Lemma 11.1.1 *Let $f : \mathbf{R}^n \rightarrow \mathbf{R}$ be a smooth function. Suppose that the Hessian of f is positive definite in a convex neighborhood $U \subset \mathbf{R}^n$. Then f has at most one local minimum in U .*

Proof: The fact that the Hessian is positive definite in U means that the restriction of f to each line segment in U is a convex function in one variable. If f has two local minima in U then let L be the line connecting these minima. The restriction of f to $L \cap U$ has two local minima and hence cannot be a convex function. This is a contradiction. ♠

Now we turn our attention to the Big and Small Theorems. Let Γ be any of the energy functions we have been considering. We have the energy map $\mathcal{E}_\Gamma : B_0 \rightarrow \mathbf{R}_+$ given by

$$\mathcal{E}_\Gamma(x_1, \dots, x_7) = \sum_{i < j} \Gamma(\Sigma^{-1}(p_i) - \Sigma^{-1}(p_j)). \quad (11.1)$$

Here we have set $p_4 = \infty$, and $p_0 = (x_1, 0)$ and $p_i = (x_{2i}, x_{2i+1})$ for $i = 1, 2, 3$. As usual Σ^{-1} is inverse stereographic projection. See Equation 7.1.

It follows from symmetry, and also from a direct calculation, that TBP is a critical point for \mathcal{E}_Γ . Assuming that Γ is one of $G_3, G_4, G_5^b, G_6, G_{10}^\sharp$, the Big Theorem now follows from Lemma 10.3.1 and from the statement that the Hessian of \mathcal{E}_Γ is positive definite at every point of B_0 . Assuming that $\Gamma = G_{10}^\sharp$, the Small Theorem follows from the Big Theorem, from Lemma 10.3.2, and from the fact that \mathcal{E}_Γ is positive definite throughout B_0 .

To finish the proof of the Big Theorem and the Small Theorem, we will show that in each case \mathcal{E}_Γ is positive definite throughout B_0 . This is the same as saying that all the eigenvalues of the Hessian H_Γ are positive at each point of B_0 . We will get this result by a two step process:

1. We give an explicit lower bound to the eigenvalues at $TBP \in B_0$.
2. We use Taylor's Theorem with Remainder to show that the positivity persists throughout B_0 .

In Step 1, the point TBP corresponds to the point

$$(1, \quad 0, -1/\sqrt{3}, \quad -1, 0, \quad 0, 1/\sqrt{3}) \in \mathbf{R}^7. \quad (11.2)$$

The spacing is supposed to indicate how the coordinates correspond to individual points. In Step 2 there are two parts. We have to compute some higher derivatives of \mathcal{E}_Γ at the TBP, and we have to bound the remainder term.

For ease of notation, we set $H_k = H_{G_k}$ and $H_5^b = H_{G_5^b}$, etc.

11.2 Lower Bounds on the Eigenvalues

Let H be a symmetric $n \times n$ real matrix. H always has an orthonormal basis of eigenvectors, and real eigenvalues. H is *positive definite* if all these eigenvalues are positive. This is equivalent to the condition that $Hv \cdot v > 0$ for all nonzero v . More generally, $Hv \cdot v \geq \lambda \|v\|$, where λ is the lowest eigenvalue of H .

Lemma 11.2.1 (Alternating Criterion) *Suppose $\chi(t)$ is the characteristic polynomial of H . Suppose that the coefficients of $P(t) = \chi(t + \lambda)$ are alternating and nontrivial. Then the lowest eigenvalue of H exceeds λ .*

Proof: An alternating polynomial has no negative roots. So, if $\chi(t + \lambda) = 0$ then $t > 0$ and $t + \lambda > \lambda$. ♠

Let us work out H_3 in detail. We first compute numerically that the lowest eigenvalue at TBP is about 14.0091. This gives something to aim for. We then compute, using Mathematica, that

$$64 \det(H_3 - (t+14)I) = 198935338432 - 21803700961600t + 2456945274144t^2 - 98631799232t^3 + 1740726332t^4 - 13792044t^5 + 49024t^6 - 64t^7$$

Since this polynomial is alternating we see (rigorously) that the lowest eigenvalue of $H_3(TBP)$ exceeds 14. Similar calculations deal with the remaining energies. Here are the results (with the first one repeated):

- the lowest eigenvalue of $H_3(TBP)$ exceeds 14.
- the lowest eigenvalue of $H_4(TBP)$ exceeds 40..
- the lowest eigenvalue of $H_5^b(TBP)$ exceeds 53.
- the lowest eigenvalue of $H_6(TBP)$ exceeds 91.
- the lowest eigenvalue of $H_{10}^\sharp(TBP)$ exceeds 1249.
- the lowest eigenvalue of $H_{10}^{\sharp\sharp}(TBP)$ exceeds 3145.

In all cases the polynomials we get have dyadic rational coefficients.

11.3 Taylor's Theorem with Remainder

Before we get to the discussion of the variation of the eigenvalues, we repackage a special case of Taylor's Theorem with Remainder. Here are some preliminary definitions.

- Let $P_0 \in \mathbf{R}^7$ be some point.
- Let B denote some cube of in-radius ϵ centered at P_0 .
- $\phi : \mathbf{R}^7 \rightarrow \mathbf{R}$ be some function.
- Let $\partial_I \phi$ be the partial derivative of ϕ w.r.t. a multi-index $I = (i_1, \dots, i_7)$.

- Let $|I| = i_1 + \cdots + i_7$. This is the *weight* of I .
- Let $I! = i_1! \cdots i_7!$.
- Let $\Delta^I = x^{i_1} \cdots x^{i_7}$. Here $\Delta = (x_1, \dots, x_7)$ is some vector.
- For each positive integer N let

$$M_N(\phi) = \sup_{|I|=N} \sup_{P \in B} |\partial_I \phi(P)|, \quad \mu_N(\phi) = \sup_{|I|=N} |\partial_I \phi(P_0)|. \quad (11.3)$$

Let U be some open neighborhood of B . Given $P \in B$, let $\Delta = P - P_0$. Taylor's Theorem with Remainder says that there is some $c \in (0, 1)$ such that

$$\phi(P) = \sum_{a=0}^N \sum_{|I|=a} \frac{|\partial_I \phi(P_0)|}{I!} \Delta^I + \sum_{|I|=N+1} \frac{\partial_I f(P_0 + c\Delta)}{I!} \Delta^I$$

Using the fact that

$$|\Delta^I| \leq \epsilon^{|I|}, \quad \sum_{|I|=m} \frac{1}{I!} = \frac{7^m}{m!},$$

and setting $N = 4$ we get

$$\sup_{P \in B} |\phi(P)| \leq |\phi(P_0)| + \sum_{j=1}^4 \frac{(7\epsilon)^j}{j!} \mu_j(\phi) + \frac{(7\epsilon)^5}{(5)!} M_5(\phi). \quad (11.4)$$

11.4 Variation of the Eigenvalues

Let H_0 be some positive definite symmetric matrix and let Δ be some other symmetric matrix of the same size. Recall various definitions of the L_2 matrix norm:

$$\|\Delta\|_2 = \sqrt{\sum_{ij} \Delta_{ij}^2} = \sqrt{\text{Trace}(\Delta \Delta^t)} = \sup_{\|v\|=1} \|\Delta v\|. \quad (11.5)$$

Lemma 11.4.1 (Variation Criterion) *Suppose that $\|\Delta\|_2 \leq \lambda$, where λ is some number less than the lowest eigenvalue of H_0 . Then $H = H_0 + \Delta$ is also positive definite.*

Proof: H is positive definite if and only if $Hv \cdot v > 0$ for every nonzero unit vector v . Let v be such a vector. Writing v in an orthonormal basis of eigenvectors we see that $H_0v \cdot v > \lambda$. Hence

$$Hv \cdot v = (H_0v + \Delta v) \cdot v \geq H_0v \cdot v - |\Delta v \cdot v| > \lambda - \|\Delta v\| \geq \lambda - \|\Delta\|_2 \geq 0.$$

This completes the proof. ♠

Let H be any of the Hessians we are considering – e.g. $H = H_3$. Define

$$F = \sqrt{\sum_{|J|=3} M_J^2}, \quad M_J = \sup_{P \in B_0} |\partial_J \mathcal{E}(P)|. \quad (11.6)$$

The sum is taken over all multi-indices J of weight 3.

Recall that $\epsilon_0 = 2^{-18}$.

Lemma 11.4.2 *Let λ be a strict lower bound to the smallest eigenvalue of $H(TBP)$. If $\sqrt{7}\epsilon_0 F \leq \lambda$. Then H is positive definite throughout B_0 .*

Proof: Let $H_0 = H(TBP)$. Let $\Delta = H - H_0$. Clearly $H = H_0 + \Delta$.

Let γ be the unit speed line segment connecting P to P_0 in \mathbf{R}^7 . Note that $\gamma \subset B_0$ and γ has length $L \leq \sqrt{7}\epsilon_0$. We set $H_L = H$ and we let H_t be the Hessian of E at the point of γ that is t units from H_0 .

We have

$$\Delta = \int_0^L D_t(H_t) dt. \quad (11.7)$$

Here D_t is the unit directional derivative of H_t along γ .

Let $(H_t)_{ij}$ denote the ij th entry of H_t . Let $(\gamma_1, \dots, \gamma_7)$ be the components of the unit vector in the direction of γ . Using the fact that $\sum_k \gamma_k^2 = 1$ and the Cauchy-Schwarz inequality, and the fact that mixed partials commute, we have

$$(D_t H_t)_{ij}^2 = \left(\sum_{k=1}^7 \gamma_k \frac{\partial}{\partial x_k} \frac{\partial^2 H_t}{\partial x_i \partial x_j} \right)^2 \leq \sum_{k=1}^7 \left(\frac{\partial^3 H_t}{\partial x_i \partial x_j \partial x_k} \right)^2. \quad (11.8)$$

Summing this inequality over i and j we get

$$\|D_t H_t\|_2^2 \leq \sum_{i,j,k} \left(\frac{\partial^3 H_t}{\partial x_i \partial x_j \partial x_k} \right)^2 \leq F^2. \quad (11.9)$$

Hence

$$\|\Delta\|_2 \leq \int_0^L \|D_t(H_t)\|_2 dt \leq LF \leq \sqrt{7}\epsilon_0 F < \lambda. \quad (11.10)$$

This lemma now follows immediately from Lemma 11.4.2. ♠

Referring to Equation 11.3, and with respect to the neighborhood B_0 , define $M_8(\mathcal{E})$ and $\mu_j(\mathcal{E})$ for $j = 4, 5, 6, 7$. Let J be any multi-index of weight 3. Using the fact that

$$\mu_j(\partial_J \mathcal{E}) \leq \mu_{j+3}(\mathcal{E}), \quad M_5(\partial_J \mathcal{E}) \leq M_8(\mathcal{E}),$$

we see that Equation 11.4 gives us the bound

$$M_J \leq |\partial_J \mathcal{E}(P_0)| + \sum_{j=1}^4 \frac{(7\epsilon_0)^j}{j!} \mu_{j+3}(\mathcal{E}) + \frac{(7\epsilon_0)^5}{5!} M_8(\mathcal{E}). \quad (11.11)$$

11.5 The Biggest Term

In this section we will prove that the last term in Equation 11.11 is at most $7^5/5!$, which is in turn less than 141. The tiny choice of $\epsilon_0 = 2^{-18}$ allows for very crude calculations.

$$f_k(a, b) = \left(4 - \|\Sigma^{-1}(a, b) - (0, 0, 1)\|^2\right)^k = 4^k \left(\frac{a^2 + b^2}{1 + a^2 + b^2}\right)^k. \quad (11.12)$$

$$\begin{aligned} g_k(a, b, c, d) &= \left(4 - \|\Sigma^{-1}(a, b) - \Sigma^{-1}(c, d)\|^2\right)^k = \\ &= 4^k \left(\frac{1 + 2ac + 2bd + (a^2 + b^2)(c^2 + d^2)}{(1 + a^2 + b^2)(1 + c^2 + d^2)}\right)^k \end{aligned} \quad (11.13)$$

Note that

$$f_k(a, b) = \lim_{c^2 + d^2 \rightarrow \infty} g_k(a, b, c, d). \quad (11.14)$$

We have

$$\begin{aligned} \mathcal{E}_k(x_1, \dots, x_7) &= f_k(x_1, 0) + f_k(x_2, x_3) + f_k(x_4, x_5) + f_k(x_6, x_7) + \\ &+ g_k(x_1, 0, x_2, x_3) + g_k(x_1, 0, x_4, x_5) + g_k(x_1, 0, x_6, x_7) + \end{aligned}$$

$$g_k(x_2, x_3, x_4, x_5) + g_k(x_2, x_3, x_6, x_7) + g_k(x_4, x_5, x_6, x_7).$$

Each variable appears in at most 4 terms, 3 of which appear in a g -function and 1 of which appears in an f -function. Hence

$$M_8(\mathcal{E}_k) \leq M_8(f_k) + 3M_8(g_k) \leq 4M_8(g_k). \quad (11.15)$$

The last inequality is a consequence of Equation 11.14 and we use it so that we can concentrate on just one of the two functions above.

Lemma 11.5.1 *When r, s, D are non-negative integers and $r + s \leq 2D$,*

$$\left| \frac{x^r y^s}{(1 + x^2 + y^2)^D} \right| < 1.$$

Proof: The quantity factors into expressions of the form $|x^\alpha y^\beta / (1 + x^2 + y^2)|$ where $\alpha + \beta \leq 2$. Such quantities are bounded above by 1. ♠

For any polynomial Π , let $|\Pi|$ denote the sum of the absolute values of the coefficients of Π . For each 8th derivative $D_I g_k$, we have

$$D_I g_k = \frac{\Pi(a, b, c, d)}{(1 + a^2 + a^2)^{k+8} (1 + c^2 + d^2)^{k+8}}, \quad (11.16)$$

Where Π_I is a polynomial of maximum (a, b) degree at most $2k + 16$ and maximum (c, d) degree at most $2k + 16$. Lemma 11.5.1 then gives

$$\sup_{(a,b,c,d) \in \mathbf{R}^4} |D_I g_k(a, b, c, d)| \leq |\Pi_I|. \quad (11.17)$$

Define

$$\Omega_k = \max_{j=1, \dots, k} M_8(g_k). \quad (11.18)$$

We compute in Mathematica that

$$\Omega_6 \leq \sup_{k=1, \dots, 6} \sup_I |\Pi_I| = 13400293856913653760 \quad (11.19)$$

The max is achieved when $k = 6$ and $I = (8, 0, 0, 0)$ or $(0, 8, 0, 0)$, etc. We also compute that

$$\Omega_{10} \leq \sup_{k=1, \dots, 10} \sup_I |\Pi_I| = 162516942801336639946752000 \quad (11.20)$$

The max is achieved when $k = 10$ and $I = (8, 0, 0, 0)$ or $(0, 8, 0, 0)$, etc.

The quantity

$$4(\Omega_{10} + 130\Omega_6) < 2^{90} \quad (11.21)$$

serves as a grand upper bound to $M_8(\mathcal{E})$ for all the energies we consider.

This leads to

$$\frac{(7 \times 2^{-18})^5}{5!} \times M_8(\mathcal{E}) < 7^5/5! < 141. \quad (11.22)$$

11.6 The Remaining Terms

In this section we show that the sum of all the terms in Equation 11.11 is less than 750.

Let

$$\mu_{j+3,k}^* = \frac{(7\epsilon_0)^j}{(j)!} \mu_{j+3}(\mathcal{E}_{G_k}) \quad (11.23)$$

We now estimate $\mu_{j,k}^*$ for $j = 4, 5, 6, 7$ and $k = 2, 3, 4, 5, 6, 10$.

The same considerations as above show that

$$\mu_{j+3,k}^* \leq \frac{(7\epsilon_0)^j}{(j)!} (\mu_{j+3}(f_k) + 3\mu_{j+3}(g_k)). \quad (11.24)$$

Here we are evaluating the $(j+3)$ rd partial at all points which arise in the TBP configuration and then taking the maximum. For instance, for g_k , one choice would be $(a, b, c, d) = (1, 0, 0, 1/\sqrt{3})$.

Again using Mathematica, we compute

$$\sup_{j=5,6,7} \sup_{k \leq 10} \mu_{j,k}^* < 1/130, \quad \sup_{k \leq 6} \mu_{4,k}^* < 4, \quad \sup_{k \leq 10} \mu_{4,k}^* < 60. \quad (11.25)$$

Therefore for any of the eneries we have, we get the bounds

$$\mu_4^*(\mathcal{E}) < 60 + 4 \times 130 = 580, \quad \mu_j^*(\mathcal{E}) < (1/130) + 130 \times (1/130) < 2. \quad (11.26)$$

The sum of all the terms in Equation 11.11 is at most

$$580 + 2 + 2 + 2 + 141 < 750. \quad (11.27)$$

11.7 The End of the Proof

For any real vector $V = (V_1, \dots, V_{343})$ define

$$\bar{V} = (|V_1| + 750, \dots, |V_{343}| + 750). \quad (11.28)$$

Given one of our energies, let V denote the vector of third partials of \mathcal{E} , evaluated at TBP and ordered (say) lexicographically. In view of the bounds in the previous section, we have $F \leq \|\bar{V}\|$. Hence

$$\sqrt{7}\epsilon_0 F \leq \sqrt{7}\epsilon_0 \|\bar{V}\|. \quad (11.29)$$

Letting V_k denote the vector corresponding to G_k , we compute

$$\sup_{k=1,\dots,6} 7\epsilon^0 \|V_k\| < 1, \quad \sup_{k=1,\dots,10} 7\epsilon_0 \|V_k\| < 60. \quad (11.30)$$

This gives us the following bounds:

- $7\epsilon_0 F_k < 1$ for $k = 3, 4, 5, 6$.
- $7\epsilon_0 F_5^b < 1 + 25 = 26$.
- $7\epsilon_0 F_{10}^\sharp < 60 + 13 + 68 = 141$.
- $7\epsilon_0 F_{10}^\sharp < 60 + 28 + 102 = 190$.

In all cases, the bound is less than the corresponding lowest eigenvalue. This completes our proof that the Hessian in all cases is positive definite throughout B_0 . Our proofs of the Big Theorem and the Small Theorem are done.

Part IV

Symmetrization

Chapter 12

Preliminaries

This chapter is something of a grab bag. The purpose is to gather together the various ingredients which go into the proofs of the Symmetrization Lemma.

12.1 Monotonicity Lemma

Lemma 12.1.1 (Monotonicity) *Let $\lambda_1, \dots, \lambda_n > 0$. If $\beta/\alpha > 1$. Then*

$$\sum_{i=1}^N \lambda_i^\beta - N \geq \frac{\beta}{\alpha} \left(\sum_{i=1}^N \lambda_i^\alpha - N \right).$$

Proof: Replacing λ_i by λ_i^α we can assume without loss of generality that $1 = \alpha < \beta$. Now we have a constrained optimization problem. We are fixing $\sum x_i$ and we want to minimize $\sum x_i^\beta$ when $\beta > 1$. An easy exercise in Lagrange multipliers shows that the min occurs when $\lambda_1 = \dots = \lambda_N$. So it suffices to consider the case $N = 1$.

Write $\epsilon = \lambda - 1$. Suppose first that $\epsilon > 0$. Then

$$\lambda^\beta - 1 = \int_{x=1}^{1+\epsilon} \beta x^{\beta-1} dx > \int_{x=1}^{1+\epsilon} \beta dx = \beta\epsilon.$$

On the other hand, if $\epsilon < 0$ then we have $1 + \epsilon \in (0, 1)$ and

$$\lambda^\beta - 1 = - \int_{1+\epsilon}^1 \beta x^{\beta-1} dx > - \int_{1+\epsilon}^1 \beta dx = \beta\epsilon.$$

That does it. ♠

12.2 A Three Step Process

Our symmetrization is a retraction from **SMALL** to **K4**. Here we describe it as a 3 step process. It is convenient to set $\Omega = \mathbf{SMALL}$. We start with the configuration X having points $(p_1, p_2, p_3, p_4) \in \Omega$.

1. We let (p'_1, p'_2, p'_3, p'_4) be the configuration which is obtained by rotating X about the origin so that p'_0 and p'_2 lie on the same horizontal line, with p'_0 lying on the right. This slight rotation does not change the energy of the configuration, but it does slightly change the domain. While $X \in \Omega$, the new configuration X' lies in a slightly modified domain Ω' which we will describe in §13.
2. Given a configuration $X' = (p'_0, p'_1, p'_2, p'_3) \in \Omega'$, there is a unique configuration $X'' = (p''_0, p''_1, p''_2, p''_3)$, invariant under reflection in the y -axis, such that p''_j and p''_j lie on the same horizontal line for $j = 0, 1, 2, 3$ and $\|p''_0 - p''_2\| = \|p'_0 - p'_2\|$. There is a slightly different domain Ω'' which contains X'' . Again, we will describe Ω'' below. This is a 6-dimensional domain.
3. Given a configuration $X'' = (p''_0, p''_1, p''_2, p''_3) \in \Omega''$ there is a unique configuration $X''' = (p'''_0, p'''_1, p'''_2, p'''_3) \in \mathbf{K4}$ such that p'''_j and p'''_j lie on the same vertical line for $j = 0, 1, 2, 3$. The configuration X''' coincides with the configuration X^* defined in the Symmetrization Lemma.

To prove the Symmetrization Lemma, it suffices to prove the following two results.

Lemma 12.2.1 $R_s(X'') \leq R_s(X')$ for all $X' \in \Omega'$ and $s > 2$, with equality iff $X' = X''$.

Lemma 12.2.2 $R_s(X''') \leq R_s(X'')$ for all $X'' \in \Omega''$ and $s \in [12, 15 + 1/2]$, with equality iff $X'' = X'''$.

12.3 Base and Bows

There are 10 bonds (i.e. distances between pairs of points) in a 5 point configuration. Ultimately, when we perform the symmetrization operations we want to see that a sum of 10 terms decreases. We find it convenient

to break this sum into 3 pieces and (for the most part) study the pieces separately. We will fix some exponent s for the power law. Given a finite list V_1, \dots, V_k of vectors, we define

$$R_s(V_1, \dots, V_k) = \sum_{i=1}^k \|V_i - V_{i-1}\|^{-s}. \quad (12.1)$$

The indices are taken mod k .

We define the *base energy* of the configuration X to be

$$A_s(X) = R_s(\hat{p}_0, \hat{p}_1, \hat{p}_2, \hat{p}_3), \quad \hat{p}_k = \Sigma^{-1}(p_k). \quad (12.2)$$

Here Σ^{-1} is stereographic projection, as in Equation 7.1.

We define the *bow energies* to be the two terms

$$B_{02,s}(X) = R_s(\hat{p}_0, \hat{p}_2, (0, 0, 1)), \quad B_{13,s}(X) = R_s(\hat{p}_1, \hat{p}_3, (0, 0, 1)). \quad (12.3)$$

We have

$$R_s(X) = A_s(X) + B_{02,s}(X) + B_{13,s}(X). \quad (12.4)$$

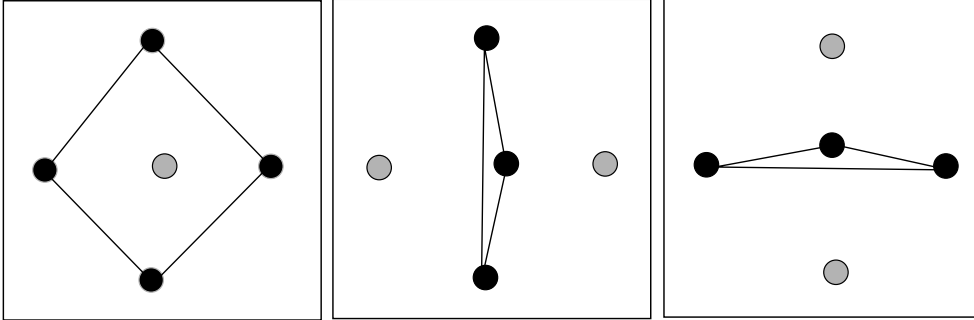


Figure 12.1: Base and Bows

Our construction above refers to Figure 12.1. The left hand side of Figure 12.1 illustrates the bonds that go into the base energy. The two figures on the right in Figure 12.1 illustrate the bonds that go into the two bow energies. If we interpret Figure 12.1 as displaying points in the plane, then the central dot is really the point at infinity.

12.4 Proof Strategy

To prove Lemma 12.2.1 we will establish the following stronger results. These results are meant to hold for all $s \geq 2$ and all $X' \in \Omega'$. (Again, we define Ω' in §13 below.)

1. $A_s(X'') \leq A_s(X')$.
2. $B_{02,s}(X'') \leq A_{02,s}(X')$ with equality iff $X' = X''$.
3. $B_{13,s}(X'') \leq A_{13,s}(X')$ with equality iff $X' = X''$.

Inequalities 2 and 3 are pretty easy; they just involve pairs of points in the plane. Inequality 1 breaks into two statements, each involving a triple of points in the plane. This inequality is considerably harder to prove, and in fact I do not know a good geometric proof. However, notice that the left hand side of the equation is just the sum of the same term 4 times. It follows from the Monotonicity Lemma that the truth of Inequality 1 at $s = 2$ implies the truth of Inequality 1 for $s > 2$. We deal with Inequality 1 at $s = 2$ with a direct calculation.

It seems that Inequalities 1-3 also work in the case of Lemma 12.2.2, except that for Inequality 1 we need to take s fairly large. I found failures for exponents as high as $s = 9$. So, in principle, we could prove Lemma 12.2.2 using a similar strategy, except that we would pick, say $s = 12$, for Inequality 1. The problem is that the high exponent leads to enormous polynomials which are too large to analyze directly. In the remote future, perhaps such a calculation would be feasible.

When trying to understand the failure of Inequality 1 for small exponents, I noticed the following: When Inequality 1 fails, Inequalities 2 and 3 hold by a wide margin. This suggested a proof strategy. To prove Lemma 12.2.2 we will exhibit constants C_0 and C_1 , which depend on the configuration and the exponent, such that

1. $A_s(X'') - A_s(X''') \geq -C_0 - C_1$ for all $s \geq 2$,
2. $B_{02,s}(X'') - B_{02,s}(X''') \geq C_0$ for all $s \in [12, 15 + 25/512]$.
3. $B_{13,s}(X'') - B_{13,s}(X''') \geq C_1$ for all $s \in [12, 15 + 25/512]$.

We get equality in the second and third cases iff $X'' = X'''$. Adding these up we get that $R_s(X'') \geq R_s(X''')$ with equality iff $X'' = X'''$.

These inequalities are supposed to hold for all $X'' \in \Omega''$, a domain we define in §13 below. Moreover, we get equality in both Inequality 2 and Inequality 3 iff $X'' = X'''$.

Remark: Inequalities 2 and 3 are rather delicate, though part of the delicacy comes from the fact that I simply squeezed the estimates as hard as I needed to squeeze them until I got them to work. One could take the upper bound to be $15 + 1/4$ in the Inequality 3 and still have it work.

In §13 we will define the domains Ω' and Ω'' and explain how we parametrize (supersets which cover) Ω' and Ω'' by affine cubes. In §14 we will prove Lemma 12.2.1 and in §15 we will prove Lemma 12.2.2.

Chapter 13

The Domains

13.1 Basic Description

In this section, we define the domains mentioned above and prove that they have the desired properties.

Recall that Ω is the domain of points p_0, p_1, p_2, p_3 such that

1. $\|p_0\| \geq \|p_k\|$ for $k = 1, 2, 3$.
2. $512p_0 \in [433, 498] \times [0, 0]$.
3. $512p_1 \in [-16, 16] \times [-464, -349]$.
4. $512p_2 \in [-498, -400] \times [0, 24]$.
5. $512p_3 \in [-16, 16] \times [349, 464]$.

Let Ω' denote the domain of points p_0, p_1, p_2, p_3 such that p'_0 and p'_2 are on the same horizontal line, and

1. $\|p'_0\| \geq \|p'_k\|$ for $k = 1, 2, 3$.
2. $512p'_0 \in [432, 498] \times [0, 16]$.
3. $512p'_1 \in [-16, 32] \times [-465, -348]$.
4. $512p'_2 \in [-498, -400] \times [0, 16]$.
5. $512p'_3 \in [-32, 16] \times [348, 465]$.

Lemma 13.1.1 *If $X \in \Omega$ then $X' \in \Omega'$.*

Proof: Rotation about the origin does not change the norms of the points, so X' satisfies Property 1. Let $\theta_0 \geq 0$ denote the angle we must rotate in order to arrange that p'_0 and p'_2 must lie on the same horizontal line. This angle maximized when p_0 is an endpoint of its segment of constraint and p_2 is one of the two upper vertices of rectangle of constraint. Not thinking too hard which of the 4 possibilities actually realizes the max, we check for all 4 pairs (p_0, p_2) that

$$\left(R_{1/34}(p_0)\right)_2 > \left(R_{1/34}(p_2)\right)_2.$$

Here $(p)_2$ denotes the second coordinate of p and R_θ denotes counterclockwise rotation by θ . From this we conclude that $\theta_0 < 1/34$. Now we need to see that if we rotate by less than $1/34$ radians, the points belong to Ω' .

Case 1: Since p_0 lies in the positive x -axis and $1 - \cos(1/34) < 1/512$, we have $p'_{01} \in [p_{01} - 1/512, p_{01}]$. Since $\sin(1/34) < 1/32$ we have $p'_{02} \in [0, 16/512]$. These give the constraints on p'_0 in Property 2.

Case 2: When we apply R_θ , the first coordinate of p_1 increases, and does so by at most $\sin(1/34) < 1/32$. This gives the constraint on p'_{11} in Property 3. Let θ_1 denote the maximum angle that p_1 can make with the y -axis. This angle is maximized when

$$p_1 = \left(\frac{\pm 16}{512}, \frac{349}{512}\right).$$

Since $\tan(1/21) > 16/349$ we conclude that $\theta_1 < 1/21$. When we rotate, p_{12} can change by at most

$$\cos\left(\frac{1}{21} + \frac{1}{34}\right) - \cos\left(\frac{1}{21}\right) < \frac{1}{512}.$$

This gives the constraints on p'_{12} in Property 3.

Case 3: We have $p'_{22} = p'_{02}$ and $p_{21} \leq p'_{21} \leq \|p_0\|$. This gives the constraints on p'_2 in Property 4.

Case 4: This follows from Case 2 and symmetry. ♠

Let Ω'' denote the domain of configurations X'' which are invariant under reflection in the y -axis and

1. $512p''_{01} \in [416, 498]$
2. $512p''_{02} \in [0, 16]$.
3. $512p''_{12} \in [-465, -348]$.
4. $512p''_{32} \in [348, 465]$.

Note that we also have

$$p''_{21} = -p''_{01}, \quad p''_{22} = p''_{02}, \quad p''_{11} = p''_{31} = 0.$$

The domain Ω'' is 4 dimensional. It follows almost immediately from the definition of the map that $X' \in \Omega'$ implies $X'' \in \Omega''$. The lower bound of 416/512 for p'_{01} comes from averaging the lower bound (432/512) for p_{01} with the lower bound (400/512) for $-p_{21}$.

Both Ω' and Ω'' are polyhedral subsets of \mathbf{R}^8 , respectively 7 and 6 dimensional. It turns out that we will really only need to make computations on certain projections of these domains. That is, we can forget about the location of one of the points at stage and concentrate on just 3 points at a time. We now describe these projections.

Let Ω'_1 denote the domain of triples (p'_0, p'_2, p'_3) arising from configurations in Ω' . In other words, we are just forgetting the location of p'_1 . Formally, Ω'_1 is the image of Ω' under the forgetful map $(\mathbf{R}^2)^4 \rightarrow (\mathbf{R}^2)^3$. We define Ω'_3 similarly, with the roles of p'_1 and p'_3 reversed. Likewise, we define Ω''_0 to be the set of triples (p''_2, p''_1, p''_3) coming from quadruples in Ω'' . Finally, we define Ω''_2 by switching the roles played by p''_0 and p''_2 . The spaces Ω'_1 and Ω'_3 are each 5 dimensional and the spaces Ω''_0 and Ω''_2 are each 4 dimensional.

13.2 Affine Cubical Coverings

In this section we describe how to parametrize supersets which contain the domains Ω'_1 , Ω'_3 , Ω''_0 , and Ω''_2 .

Let S be a subset of Euclidean space, \mathbf{R}^N . We say that an *affine cubical covering* of S if a finite collection $(Q_1, \phi_1), \dots, (Q_m, \phi_m)$ where

- $Q_k = [0, 1]^n$ for each $k = 1, \dots, m$.

- $\phi_k : Q_k \rightarrow \mathbf{R}^N$ is an affine map for each $k = 1, \dots, m$.
- S is contained in the union $\bigcup_{k=1}^m \phi_k(Q_k)$.

In all 4 cases we consider, we will have $N = 6$ and $m = 2$. When $S = \Omega'_1$ or Ω'_3 we will have $n = 5$. When $S = \Omega''_0$ or $S = \Omega''_2$ we will have $m = 4$. Our maps will have the additional virtue that they will be defined over \mathbf{Q} . This will make it so that our final calculations are integer calculations.

Suppose we have some polynomial $f : \mathbf{R}^N \rightarrow \mathbf{R}$ and we want to see that the restriction of f to S is non-negative. Our strategy is to find a cubical parametrization of S and to show that each of the functions

$$f_k = f \circ \phi_k : Q_k \rightarrow \mathbf{R} \quad (13.1)$$

is non-negative. This gives us a collection f_1, \dots, f_k which we want show are positive (respectively non-negative) on the unit cube. For this purpose we use either the positive dominance criterion or the positive dominance algorithm.

13.3 The Coverings for Lemma 1

In this section we describe the cubical coverings for Ω'_1 and Ω'_3 . In each case, we will use 2 affine cubes. Thus, we have to describe 4 affine cubes all in all.

For rational numbers $r_1 = a_1/b_1$ and $r_2 = a_2/b_2$ we define

$$[r_1, r_2, t] = r_1(1 - t) + r_2t. \quad (13.2)$$

Thus function linearly interpolates between r_1 and r_2 as t interpolates between 0 and 1. We will always use rationals of the form $p/512$, so we write $r_j = a_j/512$ and we store the information for the above function as the pair of integers (a_1, a_2) .

An array of the form

$$\begin{bmatrix} \ell_{11}, \ell_{12} \\ \vdots \\ \ell_{n1}, \ell_{n2} \end{bmatrix} \quad (13.3)$$

denotes the map

$$f(t_1, \dots, t_n) = ([\ell_{11}/512, \ell_{12}/512, t_1], \dots, [\ell_{n1}/512, \ell_{n2}/512, t_n]). \quad (13.4)$$

Here is our cubical covering for Ω'_1 .

$$\begin{bmatrix} 498 & 432 \\ 0 & 16 \\ -32 & 16 \\ 348 & 465 \\ 0 & 64 \end{bmatrix} \quad \begin{bmatrix} 432 & 416 \\ 0 & 16 \\ -32 & 16 \\ 348 & 465 \\ 0 & 64 \end{bmatrix} \quad (13.5)$$

Now we describe the meaning of the variables. We let the first coordinate of the map in Equation 13.4 be a and the second b , and so on. Then

- $p'_{01} = a + e$ and $p''_{01} = a$.
- $p'_{02} = p''_{02} = p'_{22} = p''_{22} = b$.
- $p'_{21} = -a + e$ and $p''_{21} = -a$.
- $p'_{31} = c$ and $p''_{31} = 0$.
- $p'_{32} = p''_{32} = d$.

Here $p'_0 = (p_{01}, p'_{02})$, etc. Thus, for example,

$$p'_{01} = \frac{498(1 - t_1) + 432t_1 + 64t_5}{512}, \quad t_1, t_5 \in [0, 1].$$

We describe why the cubical covering for Ω'_1 really is a covering. Here are the relevant conditions for Ω'_1 .

1. $512p'_0 \in [432, 498] \times [0, 16]$.
2. $512p'_2 \in [-498, -400] \times [0, 16]$.
3. $512p'_3 \in [-32, 16] \times [348, 465]$.
4. $p'_{22} = p'_{02}$.
5. $|p_{21}| \leq p_{01}$.

We ignore the other conditions because we only care about getting a covering.

A comparison between the terms in our matrix and the conditions for Ω'_1 reveals that the numbers are exactly the same for all the variables except p'_{01} and p'_{21} . Our parameterization requires an explanation in this case. Figure

13.1 shows a plot of the values of (p'_{01}, p'_{21}) we pick up with our parametrization. The thickly drawn rectangle is the domain of interest to us without the condition $|p_{21}| \leq p_{01}$. When we add this condition we remove the darkly shaded triangular region. The lightly shaded regions show the projections of our two affine cubes. The matrix entry 64 above is somewhat arbitrary, and making this number smaller would cause the shaded images to retract a bit. The labels in Figure 13.1 are scaled up by 512.

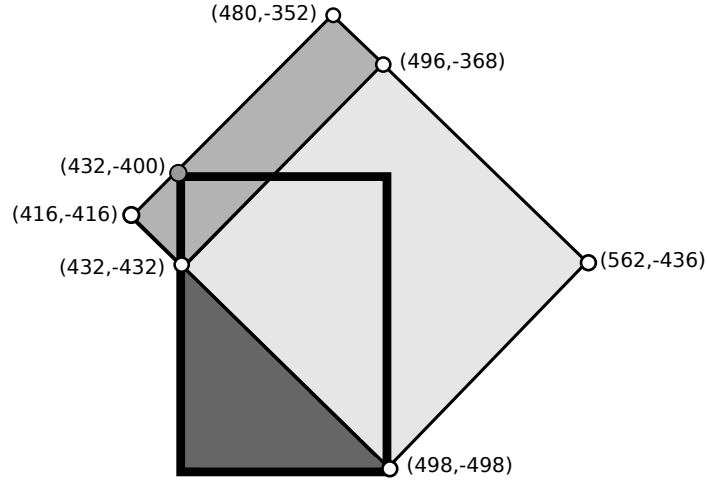


Figure 13.1: Covering Ω'_1 with affine cubes.

Remark: It would have been nicer if we could have used a cubical covering with just one affine cube. We tried this, but the resulting polynomial we got was not positive dominant. It is very likely that the positive dominance algorithm would work on this polynomial, but the polynomial is too big for the P.D.A. to run quickly enough. That is why we split things up into two cubes.

Our cubical covering for Ω'_3 is almost the same. We put in bold the numbers which have changed.

$$\begin{bmatrix} 498 & 432 \\ 0 & 16 \\ -\mathbf{16} & \mathbf{32} \\ -\mathbf{348} & -\mathbf{465} \\ 0 & 64 \end{bmatrix} \quad \begin{bmatrix} 432 & 416 \\ 0 & 16 \\ -\mathbf{16} & \mathbf{32} \\ -\mathbf{348} & -\mathbf{465} \\ 0 & 64 \end{bmatrix} \quad (13.6)$$

The variables have the same meanings, except that we use the point p_1 in place of the point p_3 . We get a covering in this case for the same reason as in the previous case.

13.4 The Coverings for Lemma 2

In this case we only need to deal with Ω_2'' . The result for Ω_0'' follows from reflection symmetry (in the y -axis.) We cover Ω_2'' using 4 affine cubes. The cubes come in pairs, and the two cubes in a pair differ only in the sign of a certain entry. The sign determines whether we are parametrizing configurations where $p_{12}'' + p_{32}''$ is positive or negative. Here are the 4 cubes:

$$\begin{bmatrix} 416 & 498 \\ 0 & 16 \\ 348 & 465 \\ 0 & \pm 24 \end{bmatrix}, \quad \begin{bmatrix} 416 & 498 \\ 0 & 16 \\ 364 & 449 \\ 0 & \pm 64 \end{bmatrix}. \quad (13.7)$$

Here are the meanings of the variables.

- $p_{01}'' = p_{01}''' = a$.
- $p_{02}'' = b$ and $p_{02}''' = 0$.
- $p_{11}'' = p_{31}'' = p_{11}''' = p_{31}''' = 0$.
- $p_{12}'' = -c + d$ and $p_{12}''' = -c$;
- $p_{32}'' = c + d$ and $p_{32}''' = c$;

Now we explain why this really gives a covering. The conditions for Ω_2'' are:

1. $512p_{01}'' \in [416, 498]$
2. $512p_{02}'' \in [0, 16]$.
3. $512p_{12}'' \in [-465, -348]$.
4. $512p_{32}'' \in [348, 465]$.

The only point about the covering that is not straightforward is why all possible coordinate pairs (p_{12}'', p_{32}'') are covered. Figure 13.2 shows the picture of how this works. The thick square is the domain of pairs we want to cover and the shaded region shows that we actually do cover. Again, the labels are scaled up by 512.

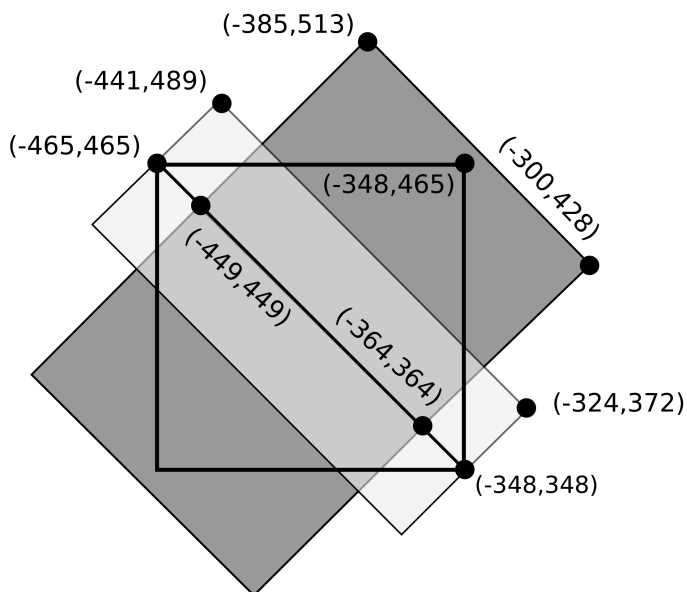


Figure 13.2: Covering Ω'' with affine cubes.

We have not labeled every point in the picture because the whole picture is invariant under the map $(x, y) \rightarrow (-y, -x)$. So, the remaining labels can be deduced from symmetry.

Chapter 14

The First Symmetrization

In this chapter we prove Theorem 12.2.1 by establishing Inequalities 1-3 listed in §12.4 concerning the configurations X' and X'' .

14.1 Inequality 1

We want to show that

$$A_s(X'') \leq A_s(X') \quad (14.1)$$

for any $X' \in \Omega'$ and any $s \geq 2$. It follows from the Monotonicity Principle that the case $s = 2$ suffices. We will split the base energy into two pieces and handle each one separately. We write, for $k = 1, 3$,

$$A_{k,2}(X') = \|\Sigma^{-1}(p_0) - \Sigma^{-1}(p_{4-k})\|^{-2} + \|\Sigma^{-1}(p_2) - \Sigma^{-1}(p_{4-k})\|^{-2}. \quad (14.2)$$

In other words, we are omitting the interactions with p'_k when we define $A_{k,2}$. The domain for $A_{k,2}$ is Ω'_k , described in §13. It clearly suffices to show that

$$A_{k,2}(X'') \leq A_{k,2}(X'), \quad k = 1, 3. \quad (14.3)$$

What we now say works for either $k = 1$ or $k = 3$. Let (ϕ_1, Q_1) and (ϕ_2, Q_2) be the cubical covering of Ω'_k described in §13.2. Again, what we say works for either of these cubes. We consider the rational function

$$\frac{P(t_1, \dots, t_5)}{Q(t_1, \dots, t_5)} = (A_{k,2}(X') - A_{k,2}(X'')) \circ \phi. \quad (14.4)$$

The polynomials are in $\mathbf{Z}[t_1, \dots, t_5]$, the expression is reduced, and Q is chosen to be positive at $(1/2, 1/2, 1/2, 1/2, 1/2)$. This determines P and Q uniquely.

We would like to show directly that this rational function is positive on $[0, 1]^5$, but we don't know how to do this directly. We take a different approach.

The two most interesting variables are t_3 and t_5 . The first of these variables controls the x -coordinate of p'_3 and the second of these controls the deviation from the points p'_0 and p'_2 being symmetrically placed about the y -axis. Setting $t_3 = t_5 = 0$ is our first symmetrization operation. We consider the Hessian with respect to these two variables:

$$H_P = \begin{bmatrix} \partial_{33}P & \partial_{35}P \\ \partial_{53}P & \partial_{55}P \end{bmatrix}. \quad (14.5)$$

Here $\partial_{33}P = \partial^2 P / \partial_{t_3}^2$, etc. This is another polynomial on $[0, 1]^5$. Here is the calculation we can make.

Theorem 14.1.1 *The polynomials $\det(H_P)$ and $\text{trace}(H_P)$ are weak positive dominant on $[0, 1]^5$. Hence, these functions are positive on $(0, 1)^5$.*

All in all, Theorem 14.1.1 involves $8 = 2 \times 2 \times 2$ polynomials. There are 2 domains, 2 affine cubes, and 2 quantities to test. Everything in sight is an integer polynomial. I did the calculation first in Mathematica, and then reprogrammed the whole thing in Java and did it again.

As is well known, a symmetric 2×2 matrix is positive definite – i.e. has positive eigenvalues – if and only if its determinant and trace are positive. Therefore, for all 4 choices, we have:

Corollary 14.1.2 *The Hessian H_P is positive definite on $(0, 1)^5$.*

Just so that we can make choices, we consider the case $k = 1$. The choice $t_3 = 2/3$ corresponds to $p''_3 = p'_3$ and the choice $t_5 = 0$ corresponds to $p'_0 = p''_0$ and $p'_2 = p''_2$. Let Π be the codimension 2 plane corresponding to $t_3 = 2/3$ and $t_5 = 0$.

Lemma 14.1.3 *P and $\partial_3 P$ and $\partial_5 P$ all vanish identically on Π .*

Proof: P vanishes by definition. $\partial_3 P$ and $\partial_5 P$ vanish by symmetry, but just to be sure we calculated symbolically that they do in fact vanish on Π . ♠

Now we come to the main point of all these calculations.

Lemma 14.1.4 $P > 0$ on $(0, 1)^5 - \Pi$.

Proof: One way to see this is to restrict P to a straight line segment L that joins an arbitrary point $\zeta \in (0, 1)^5 - \Pi$ (corresponding to a configuration $X' \neq X''$) to the point $\zeta_0 \in \Pi$ corresponding to X'' . Note that only the variables t_3 and t_5 change along L . Let f be the restriction of P to L . Since H_P is positive definite we see that f is a convex function. Moreover, our calculation above shows that $f(\zeta_0) = 0$ and $f'(\zeta_0) = 0$. But then we must have $f' > 0$ at almost all points of L . Integrating, we get $f(\zeta) > 0$. ♠

Remember that P is the numerator of a rational function P/Q .

Lemma 14.1.5 $Q > 0$ on $(0, 1)^5 - \Pi$.

Proof: We sample one point in the connected domain $(0, 1)^5 - \Pi$ to check that $Q > 0$ somewhere in this set. If Q ever vanishes, then the fact that $P > 0$ would cause our well-defined rational function to blow up. This does not happen because the rational function is well defined. Hence $Q > 0$ everywhere on $(0, 1)^5 - \Pi$. ♠

Combining the last two lemmas, we see that the function in Equation 14.4 is positive on $(0, 1)^5 - \Pi$. But then, by definition

$$A_{k,2}(X'') < A_{1,2}(X')$$

when $X' \neq X''$. This is what we wanted to prove in case $k = 1$.

The case $k = 3$ has the same proof except that we take $t_3 = 1/3$ rather than $t_3 = 2/3$,

This completes the proof of Inequality 1.

14.2 Inequality 2

In this section we prove Inequality 2. That is,

$$B_{02,s}(X'') \leq B_{02,s}(X'), \quad \forall s \geq 2, \quad (14.6)$$

with equality iff the relevant points coincide.

The relevant points here are

$$p'_0 = (x + d, y), \quad p'_2 = (-x + d, y), \quad p''_0 = (x, y), \quad p''_2 = (-x, y). \quad (14.7)$$

The way our proof works is that we further break the 3-term expression for the bow energy into a sum of an expression with one term and an expression with two terms. We show that separately these sums decrease when $s = 2$, and then we appeal the Monotonicity Lemma proved in the previous chapter to take care of the case $s > 2$. Along the way, we reveal a beautiful formula. Here are the details.

We compute

$$\|\Sigma^{-1}(p'_0) - \Sigma^{-1}(p'_2)\|^{-2} - \|\Sigma^{-1}(p''_0) - \Sigma^{-1}(p''_2)\|^{-2} = \frac{d^2}{x^2} \times (2 + d^2 - 2x^2 + 2y^2).$$

Since $2 - 2x^2 + 2y^2 > 0$ we see that this expression is always positive. This shows that one of the three terms in the bow energy decreases, when $s = 2$. When $s > 2$ we get the same energy decrease, by the Monotonicity Lemma from the previous chapter.

For the other two terms, we note the following beautiful formula.

$$\begin{aligned} & \left(\|\Sigma^{-1}(p'_0) - (0, 0, 1)\|^{-2} + \|\Sigma^{-1}(p'_2) - (0, 0, 1)\|^{-2} \right) - \\ & \left(\|\Sigma^{-1}(p''_0) - (0, 0, 1)\|^{-2} + \|\Sigma^{-1}(p''_2) - (0, 0, 1)\|^{-2} \right) = \frac{d^2}{2}, \end{aligned} \quad (14.8)$$

This holds identically for all choices of x, y, d . Thus, the sum of the other two terms decreases at the exponent $s = 2$. But then the Monotonicity Lemma gives the same result for all $s > 2$. with equality iff $d = 0$.

14.3 Results about Triangles

Before we prove Inequality 3, we need a technical lemma about triangles. The first two results in this section are surely known, but I don't have a reference. Lacking a reference, I will prove them from scratch. The third result is probably too obscure to be known, but we need it for Inequality 3. The sole purpose of this section is to prove the technical result at the end. This result feeds into the proof of Lemma 14.4.2 and is not used elsewhere.

Let C be the unit circle and let a, b be distinct points in C . Let \widehat{C} denote one of the two arcs bounded by a and b . For convenience we will draw

pictures in the case when \widehat{C} is more than a semicircle, but the result works in all cases.

Lemma 14.3.1 *The function $f(c) = \|a - c\| + \|b - c\|$ has a unique critical point at the midpoint of \widehat{C} , and this point is a maximum. In other words, the function $c \rightarrow f(c)$ increases as c moves from a to the midpoint of \widehat{C} .*

Proof: Our proof refers to Figure 14.1 below. By symmetry it suffices to consider the case when c is closer to a . We will show that $f(c)$ increases as c moves further away from a . Let V_a and V_b be unit vectors which are based at c and point from a to c and b to c respectively. Let T_c denote the tangent vector to C at c . The desired monotonicity follows from the claim that

$$T \cdot (V_a + V_b) > 0. \quad (14.9)$$

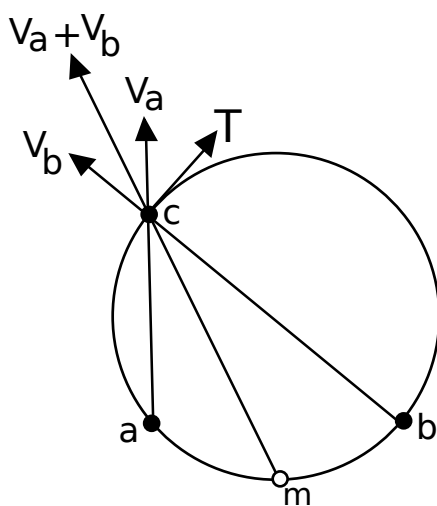


Figure 14.1: The Circle Result

Let m be the midpoint of the other arc of C bounded by a and b . The vector $V_a + V_b$ makes an equal angle with both V_a and V_b . But then, by a familiar theorem from high school geometry, $V_a + V_b$ is parallel to the ray connecting the point m to v . But this ray obviously has positive dot product with T . The result follows immediately. ♠

Here is a corollary of the preceding result.

Lemma 14.3.2 *For any $s > 0$ the function $f(c) = \|a - c\|^{-s} + \|b - c\|^{-s}$ has a unique critical point at the midpoint of \widehat{C} , and this point is a minimum. In other words, the function $c \rightarrow f(c)$ decreases as c moves from a to the midpoint of \widehat{C} .*

Proof: The key property we use is that the power law functions are convex decreasing. We use the same set up as in the previous lemma. Suppose we move c to a nearby location c' further away from a . There are constants $\epsilon_a > 0$ and ϵ_b such that

$$\|a - c'\| = \|a - c\| + \epsilon_a, \quad \|b - c'\| = \|b - c\| - \epsilon_b. \quad (14.10)$$

By inspection we have $\epsilon_a > 0$. If $\epsilon_b < 0$ then both distances have increased and the result of this lemma is obvious. So, we just have to worry about the case when $\epsilon_b > 0$. By the previous result, we have $\epsilon_a > \epsilon_b$.

We set $\epsilon = \epsilon_a$. It now follows from monotonicity that

$$f(c') < (\|a - c\| + \epsilon)^{-s} + (\|b - c\| - \epsilon)^{-s}. \quad (14.11)$$

We also have $\|a - c\| < \|b - c\|$. Finally, we know that the power function $R_s(x) = x^{-s}$ is convex decreasing. This is to say that the derivative of R_s is negative and increasing. Hence, when ϵ is small, the first term on the right side of Equation 14.11 decreases more quickly than the second term increases. This gives $f(c') < f(c)$ for ϵ sufficiently small. But then this is true independent of ϵ and remains true until $\|a - c\| = \|b - c\|$. ♠

Now we come to the useful technical lemma. As usual, let Σ^{-1} be inverse stereographic projection. Call the pair of points $((0, t), (0, u))$ a *good pair* if $-t, u < \sqrt{3}/3$. The significance of the lower bound is that the points $\Sigma^{-1}(0, \pm 1/\sqrt{3})$ form an equilateral triangle with $(0, 0, 1)$.

Let $\mathcal{E}_s(t, u)$ be the bow energy associated to this pair. That is

$$\mathcal{E}_s(t, u) = \frac{\|\Sigma^{-1}(0, -t) - (0, 0, 1)\|^{-s} + \|\Sigma^{-1}(0, u) - (0, 0, 1)\|^{-s} + \|\Sigma^{-1}(0, -t) - \Sigma^{-1}(0, u)\|^{-s}}{3} \quad (14.12)$$

For later reference, we will call the three summands *term 1*, *term 2*, and *term 3*.

Lemma 14.3.3 *If (u^*, t^*) is also a good pair with $(-t^*) \geq (-t)$ and $u^* \geq u$ then $\mathcal{E}_s(t^*, u^*) \geq \mathcal{E}_s(t, u)$.*

Proof: Since we can switch the roles of t and u , it suffices to consider the case when $t = t^*$ and just u moves. Suppressing the dependence on s , we consider the function

$$\phi(u^*) = \mathcal{E}_s(t, u^*). \quad (14.13)$$

We will look at the function geometrically. By symmetry, this function has an extremum when $U = \Sigma^{-1}(0, u^*)$ forms an isosceles triangle with the points $N = (0, 0, 1)$ and $T = \Sigma^{-1}(0 - t)$. Figure 14.2 shows the situation.

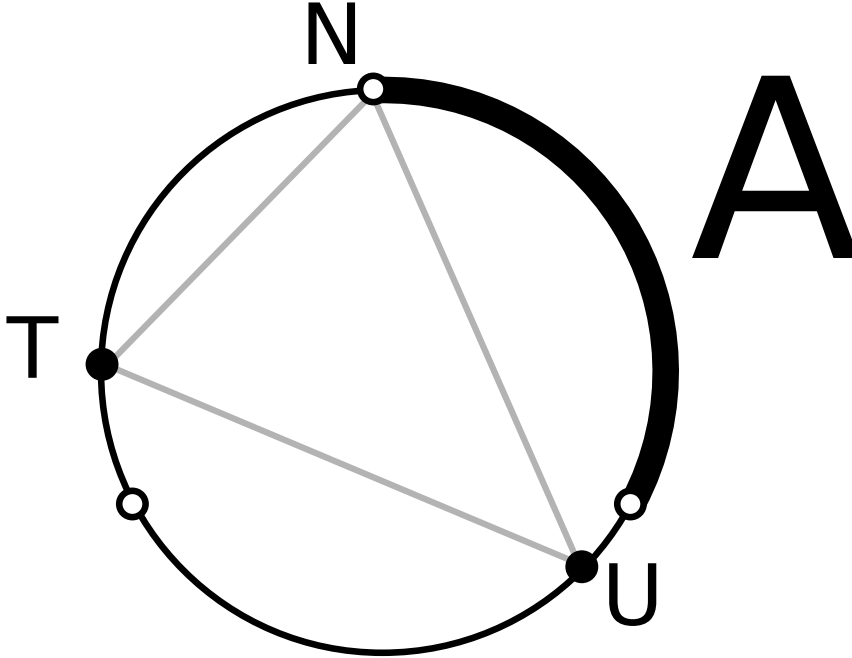


Figure 14.2: Points on the a great circle

The thickened arc A indicates the allowable locations U when (t^*, u^*) is a good pair. Significantly, the apex of the isosceles triangle we have been talking about lies outside A when we have a good pair. By Lemma 14.3.2, the function ϕ is monotone increasing on the arc A as we move away from the point U to the point N . This is what we wanted to prove. ♠

14.4 Inequality 3

In this section we prove Inequality 3. That is,

$$B_{13,s}(X'') \leq B_{13,s}(X'), \quad \forall s \geq 2, \quad (14.14)$$

with equality iff the relevant points coincide.

The relevant points here are

$$\begin{aligned} p'_1 &= (x_1, -y + d), & p'_3 &= (x_3, y + d), \\ p''_1 &= (0, -y + d), & p''_3 &= (0, y + d). \end{aligned} \quad (14.15)$$

To get from (p'_1, p'_3) to (p''_1, p''_3) we are just sliding these points along horizontal lines until they hit the y -axis. Algebraically, this sliding amounts to setting $x_1 = x_3 = 0$.

Reflecting in the y -axis, we can assume without loss of generality that $x_1 \geq 0$. Suppose that $x_3 > 0$. What this means is that the points \hat{p}_1 and \hat{p}_3 lie in the same hemisphere on S^2 . We can increase their chordal distance by reflecting one of the points in the boundary of this hemisphere. In other words, if we replace x_3 by $-x_3$ then we decrease $A_{s,13}(X')$. So, it suffices to consider the case when $x_1 \geq 0$ and $x_3 \leq 0$. Here is a critical special case.

Lemma 14.4.1 *Equation 14.14 holds when $x_1 = 0$ and $x_3 < 0$.*

Proof: We need to analyze what happens to the 3 terms in Equation 14.12. We have $p'_1 = p''_1$ so

$$\|\Sigma^{-1}(p'_1) - (0, 0, 1)\|^{-s} = \|\Sigma^{-1}(p''_1) - (0, 0, 1)\|^{-s}.$$

This takes care of term 1.

Now we deal with term 2. Note that p''_3 is closer to the origin than p'_3 so $\Sigma^{-1}(p''_3)$ is farther from $(0, 0, 1)$ than is $\Sigma^{-1}(p'_3)$. Hence

$$\|\Sigma^{-1}(p''_3) - (0, 0, 1)\|^{-s} < \|\Sigma^{-1}(p'_3) - (0, 0, 1)\|^{-s}. \quad (14.16)$$

Hence term 2 decreases.

Now we deal with term 3. By the Monotonicity Lemma, it suffices to consider the case when $s = 2$. Holding y_1 and y_3 fixed and setting $x_3 = x$, we compute

$$\|\Sigma^{-1}(0, y_1), \Sigma^{-1}(x, y_3)\|^{-2} = \frac{(1 + y_1^2)u(x)}{4}, \quad u(x) = \frac{x^2 + (y_3^2 + 1)}{x^2 + (y_1 - y_3)^2}.$$

To finish the proof it suffices to prove that $u(x)$ takes its min at $x = 0$. Once we know this, we see that term 3 decreases when $s = 2$. The Monotonicity Lemma then implies that this happens for $s > 2$.

To analyze u , we write

$$u(x) = \frac{x^2 + C_1}{x^2 + C_2}, \quad C_1 = y_3^2 + 1, \quad C_2 = (y_1 - y_3)^2. \quad (14.17)$$

This function has a unique extremum at 0. To find out if this extremum is a min or a max, we take the second derivative:

$$\frac{d^2u}{dx^2} = \frac{2(C_2 - C_1)(C_2 - 3x^2)}{(C_2 + x^2)^2}. \quad (14.18)$$

We are hoping a min, so we want to verify that the second derivative is positive. The denominator is obviously positive.

To show the positivity of the first factor in the numerator, we just use $y_1 < -\sqrt{3}/3$ and $y_3 > \sqrt{3}/3$. This gives

$$y_1^2 > 1/3, \quad -2y_1y_3 > 2/3, \quad (y_1 - y_3)^2 > 4/3.$$

Hence

$$C_2 - C_1 = -1 + y_1^2 - 2y_1y_3 > -1 + (1/3) + (2/3) = 0. \quad (14.19)$$

$$C_2 - 3x^2 > (2\sqrt{3}/3)^2 - 3(1/16)^2 > (4/3) - (1/3) = 1. \quad (14.20)$$

This takes care of term 3.

In summary, term 1 in Equation 14.12 does not change, and (unless we have equality in the configurations) term 2 decreases and term 3 decreases. ♠

Remark: It is easy to make a sign error in a calculation like the one just done. So, to be sure, I plotted the graph of the function $u(x)$ for random choices of y_1 and y_3 and say that it had a local min at $x = 0$.

The case when $x_1 > 0$ and $x_3 = 0$ has the same treatment as the case just considered, and indeed follows from this case and symmetry. So, we just have to consider the case when $x_1 > 0$ and $x_3 < 0$. In this case, the idea is to tilt the plane and see what happens. Since there are no counterexamples when one of x_1 or x_3 equals 0, the following result finishes our proof.

Lemma 14.4.2 *If Equation 14.14 has a counterexample with $x_1 > 0$ and $x_3 < 0$ it also has a counterexample when one of x_1 or x_3 equals 0.*

Proof: Suppressing the exponent s let $B(a, b)$ be the bow energy associated to a pair of points $(a, b \in \mathbf{R}^2)$.

Imagine that we had a counterexample to Equation 14.14 with $x_1 > 0$ and $x_3 < 0$. This would mean that

$$B(p'_1, p'_3) < B(p''_1, p''_3).$$

Let q'_1 and q'_3 denote the points obtained by rotating the plane about the origin until one of the points p'_1 or p'_3 hits the y -axis. By rotational symmetry, we have

$$B(q'_1, q'_3) = B(p'_1, p'_3).$$

As we rotate, the vertical distance from our points to the x -axis increases, because they are “rising up” along circles centered at the origin. This means that both q''_1 and q''_3 are further from the origin respectively than p''_1 and p''_3 .

Note that both pairs (p''_1, p''_3) and (q''_1, q''_3) are good pairs in the sense of Lemma 14.3.3. Therefore, by Lemma 14.3.3,

$$B(q''_1, q''_3) > B(p''_1, p''_3).$$

Stringing together all our inequalities, we have

$$B(q''_1, q''_3) > B(p''_1, p''_3) > B(p'_1, p'_3) = B(q'_1, q'_3).$$

We have constructed a counterexample, namely the pair (q'_1, q'_3) in which one of the two points lies on the y -axis. That is what we wanted to do. ♠

Having established Inequalities 1, 2, and 3, our proof of Theorem 12.2.1 is done.

Chapter 15

The Second Symmetrization

In this chapter we prove Theorem 12.2.2 by establishing Inequalities 1-3 listed in §12.4 concerning the configurations X'' and X''' . For reference, here are those inequalities again.

1. $A_s(X'') - A_s(X''') \geq -C_0 - C_1$ for all $s \geq 2$,
2. $B_{02,s}(X'') - B_{02,s}(X''') \geq C_0$ for all $s \in [12, 15 + 25/512]$.
3. $B_{13,s}(X'') - B_{13,s}(X''') \geq C_1$ for all $s \in [12, 15 + 25/512]$.

The constants C_0 and C_1 depend on the parameter s and on the configuration X'' . As we mentioned in §12.2.2 we will see that we get equality in the second and third cases iff $X'' = X'''$. Adding up the inequalities, we get that $R_s(X'') \geq R_s(X''')$ with equality iff $X'' = X'''$.

Our argument for Inequality 3 is quite delicate. At the end of the chapter we will explain a more complicated but more robust argument.

15.1 Inequality 1

Figure 15.1 refers to the construction in §13.4, and explains the geometrical meaning of the variables a, b, c, d discussed there.

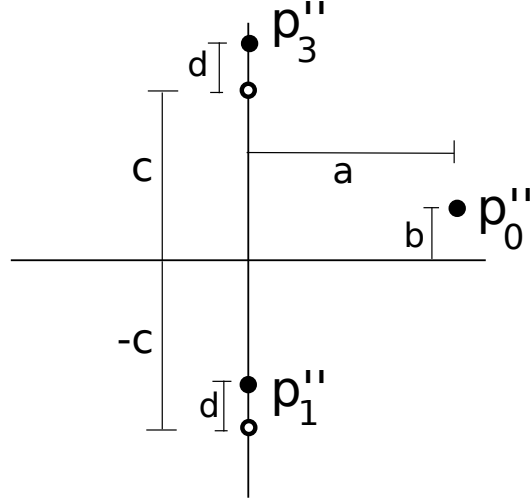


Figure 15.1: The Meaning of the Cubical Parametrization

The variables in the polynomial produced by the cubical parametrization are $t_1, t_2, t_3, t_4 \in [0, 1]$. The variables a, b, c, d are linear functions in the variables t_1, t_2, t_3, t_4 .

In this chapter we are interested in studying what happens when we hold a and c fixed and vary b and d . The point is that our symmetrization operation here only changes b and d . Suppressing the values of a and c , we introduce the notation

$$(j, k, s)_{b,d} = \|\Sigma_{-1}(p_j'') - \Sigma^{-1}(p_k'')\|^{-s}. \quad (15.1)$$

Here Σ^{-1} is inverse stereographic projection. When $b = d = 0$ we write $(j, k, s) = (j, k, s)_{00}$ for ease of notation. We are interested in proving

$$(01, s)_{bd} + (03, s)_{bd} \geq (01, s) + (03, s) \quad \forall s \in [12, 16], \quad (15.2)$$

when the variables are in their intended range. As we mentioned in §12.4, this calculation is too big for us. We have to work with a smaller exponent and claw our way back up.

We have already discussed how we need to incorporate the bow energies into our estimate to make it work. In practice we give a little “positive kick” to our polynomial and then justify that the positive kick is smaller than what we get from taking into account the bow energies. Here goes.

For each $k = 1, 2, 3, 4$, let ϕ_k be the map in our cubical covering of Ω_2'' described in §13.4. Define

$$\frac{P_k(t_1, t_2, t_3, t_4)}{Q_k(t_1, t_2, t_3, t_4)} = \left((01, 2)_{bd} + (03, 2)_{bd} - (01, 2) - (03, 2) + \frac{b^2}{94} + \frac{d^2}{13} \right) \circ \phi_k. \quad (15.3)$$

The polynomials are in $\mathbf{Z}[t_1, t_2, t_3, t_4]$, the expression is reduced, and Q is chosen to be positive at $(1/2, 1/2, 1/2, 1/2)$. This determines P and Q uniquely.

As in §14.1, we introduce the Hessians:

$$H_k = \begin{bmatrix} \partial_{22}P_k & \partial_{24}P_k \\ \partial_{42}P_k & \partial_{44}P_k \end{bmatrix}, \quad \partial_{ij}P_j = \frac{\partial^2 P_k}{\partial t_i \partial t_j}. \quad (15.4)$$

Theorem 15.1.1 *For each $k = 1, 2, 3, 4$, the functions $\det H_k$ and $\text{trace} H_k$ are weak positive dominant on $[0, 1]^4$.*

Proof: We do the calculation in Java and it works. As a sanity check, we also compute the polynomials in Mathematica and check that the two versions match. ♠

Remark: The integers 13 and 94 in our equation are the largest integers which work. I went as far as checking that some of the positive dominance fails with $13 + \frac{1}{4}$ in place of 13. The estimate of 94 is extremely loose for our purposes but the estimate of 13 is quite tight. Were we forced to use an integer less than 13 we would have to really scramble to finish our proof. See the discussion in §15.5.

Corollary 15.1.2 $P_k > 0$ on $(0, 1)^4$.

Proof: This has the same proof as Lemma 14.1.4 once we verify the same facts. It follows from symmetry (and we do a direct calculation) that $\partial_2 P_k$ and $\partial_4 P_k$ vanish when $t_2 = t_4 = 0$. We choose any point $(t_1, t_2, t_3, t_4) \in (0, 1)^4$ and join it by a line segment L to $(t_1, 0, t_3, 0)$. The two functions $\det H_k$ and $\text{trace} H_k$ are positive definition everywhere except perhaps the endpoint of L . Hence H_k is positive definite along L except perhaps at the endpoint. Hence, the restriction of P_k to L is strictly convex, except perhaps at the endpoint. The vanishing of the first partials combines with the convexity, to imply that $(t_1, 0, t_3, 0)$ is the global minimum of P_k on L . ♠

Corollary 15.1.3

$$\Pi = (01, 2)_{bd} + (03, 2)_{bd} - (01, 2) - (03, 2) + \frac{b^2}{94} + \frac{d^2}{13} \geq 0$$

on Ω_2'' .

Proof: The function P_k is positive on a connected and open dense subset of $[0, 1]^4$ and the function P_k/Q_k is well-defined and finite on $[0, 1]^4$. Hence Q_k does not vanish on a connected and open dense subset of $[0, 1]^4$. But we know that $Q_k > 0$ for the point $(1/2, 1/2, 1/2, 1/2)$. Hence $Q_k > 0$ on an open dense subset of $[0, 1]^4$. Hence $P_k/Q_k > 0$ on an open dense subset of $[0, 1]^4$. Hence $P_k/Q_k \geq 0$ on $[0, 1]^4$. But the union of our affine cubes contains Ω_2'' . Hence, by definition, $\Pi \geq 0$ on Ω_2'' . ♠

Now we know that

$$\left((01, 2)_{bd} + (03, 2)_{bd} \right) - \left((01, 2) + (03, 2) \right) \geq -\left(\frac{b^2}{94} + \frac{d^2}{13} \right). \quad (15.5)$$

Dividing both sides of Equation 15.5 by $(01, 2)$,

$$\frac{(01, 2)_{bd}}{(01, 2)} + \frac{(03, 2)_{bd}}{(01, 2)} - 2 \geq -\left(\frac{b^2}{94} + \frac{d^2}{13} \right)(01, -2). \quad (15.6)$$

Applying the Monotonicity Lemma in the case $N = 2$ and $\alpha = 2$ and $b = s$ we have

$$\frac{(01, s)_{bd}}{(01, s)} + \frac{(03, s)_{bd}}{(01, s)} - 2 \geq -\frac{s}{2} \left(\frac{b^2}{94} + \frac{d^2}{13} \right)(01, -2). \quad (15.7)$$

Clearing denominators, and using the fact that $(01, s) = (03, s)$ we have

$$(01, s)_{bd} + (03, s)_{bd} - (01, s) - (03, s) \geq -\frac{s}{2} \left(\frac{b^2}{94} + \frac{d^2}{13} \right)(01, s - 2). \quad (15.8)$$

The calculation above deals with half the bonds comprising the base energy. By reflection symmetry, we get the same result for the other two bonds. Hence

$$A_s(X'') - A_s(X''') \geq -s \left(\frac{b^2}{94} + \frac{d^2}{13} \right)(01, s - 2). \quad (15.9)$$

We set

$$C_0 = \frac{b^2 s}{94}(01, s - 2), \quad C_1 = \frac{d^2 s}{13}(01, s - 2). \quad (15.10)$$

This gives us Inequality 1.

15.2 Inequality 2

We fix the variables a, c, d and study the dependence of our points on the variable b . We define

$$(k, s)_b = \|\Sigma^{-1}(p_k'') - (0, 0, 1)\|^{-s} \quad (15.11)$$

We also set $(k, s) = (k, s)_0$. A direct calculation shows

$$(0, 2)_b - (0, 2) = \frac{b^2}{4}. \quad (15.12)$$

Therefore

$$\frac{(0, 2)_b}{(0, 2)} - 1 = \frac{b^2}{4}(0, -2). \quad (15.13)$$

By the Monotonicity Lemma,

$$\frac{(0, s)_b}{(0, s)} - 1 \geq \frac{b^2 s}{8}(0, -2). \quad (15.14)$$

Hence

$$(0, s)_b - (0, s) \geq \frac{b^2 s}{8}(0, s - 2). \quad (15.15)$$

Since $(0, s)_b = (2, s)_b$ for all b , we have

$$(2, s)_b - (2, s) \geq \frac{b^2 s}{8}(0, s - 2). \quad (15.16)$$

Finally a direct calculation shows that

$$(02, s)_b - (02, s) = \frac{b^s}{16x^s} \times (2 + 2x^2 + b^2)^{s/2} \geq 0. \quad (15.17)$$

Adding these last 3 equations together, we get

$$B_{02,s}(X'') - B_{02,s}(X''') \geq \frac{b^2 s}{4}(0, s - 2). \quad (15.18)$$

We want to show that the expression on the left hand side of this equation is greater than or equal to C_0 , with equality if and only if $b = 0$. Comparing this equation with Inequality 1, we see that it suffices to prove

$$\frac{(0, s - 2)}{(01, s - 2)} > \frac{4}{94}. \quad (15.19)$$

Setting $p_0''' = (x, 0)$ and $p_3''' = (0, y)$, we compute that

$$\frac{(0, s-2)}{(01, s-2)} = \left(\frac{x^2 + y^2}{1 + y^2} \right)^{s/2-1}. \quad (15.20)$$

Since $x < 1$, the right hand side decreases when we decrease x and y and increase s . Therefore, the minimum occurs at

$$x = \frac{416}{512} \quad y = \frac{348}{512} \quad s = 15 + \frac{25}{512} \quad (15.21)$$

When we plug in these values and use Mathematica's ability to compute numerical approximations to rational powers of rational numbers, we get

$$0.1779791356057403219045637477452967293894273925... \quad (15.22)$$

which is certainly greater than $4/94$. This establishes Inequality 2.

15.3 Improved Monotonicity

We plan to carry out the same analysis for Inequality 3 that we carried out for Inequality 2, but the estimate

$$\frac{(1, s)_b}{(1, s)} + \frac{(3, s)_b}{(3, s)} - 2 \geq \frac{s}{2} \left(\frac{(1, 2)_b}{(1, 2)} + \frac{(3, 2)_b}{(3, 2)} - 2 \right)$$

coming from the Monotonicity Lemma is not good enough. In this section we improve the estimate by roughly a factor of 4 within a smaller range of exponents.

Lemma 15.3.1 *For $s \geq 12$ we have*

$$\frac{(1, s)_b}{(1, s)} + \frac{(3, s)_b}{(3, s)} - 2 \geq 2s \left(\frac{(1, 2)_b}{(1, 2)} + \frac{(3, 2)_b}{(3, 2)} - 2 \right)$$

Proof: Suppose we know this result for $s = 12$. Then for $s > 12$ we have

$$\frac{(1, s)_b}{(1, s)} + \frac{(3, s)_b}{(3, s)} - 2 \geq \frac{s}{12} \left(\frac{(1, 12)_b}{(1, 12)} + \frac{(3, 12)_b}{(3, 12)} - 2 \right) \geq$$

$$\frac{s}{12} \times 24 \times \left(\frac{(1,2)_b}{(1,2)} + \frac{(3,2)_b}{(3,2)} - 2 \right) = 2s \left(\frac{(1,2)_b}{(1,2)} + \frac{(3,2)_b}{(3,2)} - 2 \right).$$

The first inequality is the Monotonicity Lemma. So, we just have to prove our result when $s = 12$.

Setting $p_1'' = (0, -y + d)$ and $p_3'' = (0, y + d)$ we compute that

$$\frac{(1,12)_b}{(1,12)} + \frac{(3,12)_b}{(3,12)} - 2 = \frac{P(y,d)}{Q(y,d)} \left(\frac{(1,2)_b}{(1,2)} + \frac{(3,2)_b}{(3,2)} - 2 \right),$$

Where P and Q are entirely positive polynomials. The positivity is a piece of good luck. Specifically, we have

$$\begin{aligned} P = & 6 + 15d^2 + 20d^4 + 15d^6 + 6d^8 + d^{10} + 90y^2 + 300d^2y^2 + 420d^4y^2 + \\ & 270d^6y^2 + 66d^8y^2 + 300y^4 + 1050d^2y^4 + 1260d^4y^4 + 495d^6y^4 + \\ & 420y^6 + 1260d^2y^6 + 924d^4y^6 + 270y^8 + 495d^2y^8 + 66y^{10}. \end{aligned} \quad (15.23)$$

$$Q = (1 + y^2)^5 \quad (15.24)$$

Since d does not appear in the denominator, the ratio above is minimized when $d = 0$. When $d = 0$ the expression simplifies to

$$\frac{P}{Q} = \frac{6 + 66y^2}{1 + y^2}. \quad (15.25)$$

This function is monotone increasing for $y > 0$ and exceeds 24 when we plug in $y = 87/128$, the minimum allowable value for points in Ω'' . ♠

15.4 Inequality 3

We take $s \geq 12$. A direct calculation shows

$$(1,2)_d + (3,2)_d - (1,2) - (3,2) = \frac{d^2}{2}. \quad (15.26)$$

Using the fact that $(1, 2) = (3, 2)$, we have

$$\frac{(1, 2)_d}{(1, 2)} + \frac{(3, 2)_d}{(1, 2)} - 2 = \frac{d^2}{2}(1, -2). \quad (15.27)$$

When $s \geq 12$ Lemma 15.3.1 gives

$$\frac{(1, s)_d}{(1, s)} + \frac{(3, s)_d}{(1, s)} - 2 \geq d^2 s \times (1, -2). \quad (15.28)$$

Hence

$$(1, s)_d + (3, s)_d - (1, s) - (3, s) \geq d^2 s \times (1, s - 2). \quad (15.29)$$

Moreover, just as in Equation 15.17, a direct calculation shows

$$(13, s)_d - (13, s) = \frac{d^s}{16y^s} \times (2 - 2y^2 + d^2)^{s/2} \geq 0. \quad (15.30)$$

This time we have to use the fact that $y^2 < 1$. Adding the last two equations, we get

$$B_{13,s}(X'') - B_{13,s}(X''') \geq d^2 s \times (1, s - 2). \quad (15.31)$$

We want to show that the expression on the left hand side of this equation is greater than or equal to C_1 , with equality if and only if $d = 0$. Comparing this equation with Inequality 1, we see that it suffices to prove

$$\frac{(1, s - 2)}{(01, s - 2)} > \frac{1}{13} = 0.0761239... \quad (15.32)$$

Using the same parameters as in Inequality 2, we compute

$$\frac{(1, s - 2)}{(01, s - 2)} = \left(\frac{x^2 + y^2}{1 + x^2} \right)^{s/2 - 1}. \quad (15.33)$$

The minimum value occurs for the same values as in Inequality 2, namely

$$x = \frac{416}{512} \quad y = \frac{348}{512} \quad s = 15 + 25/512 \quad (15.34)$$

When we plug in these values, and use Mathematica, we get

$$0.0776538129762009654816474958983834370212062922715..., \quad (15.35)$$

which exceeds $1/13$. This establishes Inequality 3.

Our proof of Lemma 12.2.2 is done. This completes the proof of the Symmetrization Lemma.

15.5 Discussion

Our proof of Inequality 2 left a wide margin for error. For instance, it works for all $s \in [2, 16]$ and indeed for s considerably larger than 16. However, our proof of Inequality 3 seems to be quite a close call. This is somewhat of an illusion. As we mentioned already, we just squeezed the estimates until they delivered what we needed. Here we sketch an alternate proof of Inequality 3 that leaves more room for relaxation, and incidentally covers all exponents in $[12, 16]$.

We consider the affine cubes given by the arrays.

$$\begin{bmatrix} 416 & \mathbf{470} \\ 0 & 16 \\ 348 & 465 \\ 0 & \pm 24 \end{bmatrix}, \quad \begin{bmatrix} 416 & \mathbf{470} \\ 0 & 16 \\ 364 & 449 \\ 0 & \pm 64 \end{bmatrix}. \quad (15.36)$$

The bolded numbers have changed from 498 to 470 and the rest is the same as in §13.4. We redefine the functions in Equation 15.3 but instead we use the functions in §13.4.

$$\frac{P_k(t_1, t_2, t_3, t_4)}{Q_k(t_1, t_2, t_3, t_4)} = \left((01, 2)_{bd} + (03, 2)_{bd} - (01, 2) - (03, 2) + \frac{b^2}{94} + \frac{d^2}{\mathbf{39}} \right) \circ \phi_k. \quad (15.37)$$

All the positive dominance goes through, and so in this region we could replace the constant C_1 from Inequality 1 with $C_1/3$. Thus, we would have an easy time eliminating the configurations covered by the above affine cubes.

What remains is a subdomain of Ω_2'' consisting of points (p_1, p_2, p_3, p_4) where $p_{01} \geq 470/512$. Since we have already taken care of the rest of Ω_2'' , for the remaining points we can instead use the bound on Equation 15.33 using the points

$$x = \frac{470}{512} \quad y = \frac{348}{512} \quad s = 15 + 25/512 \quad (15.38)$$

When we plug in these numbers we get a much better lower bound of 0.105... in place of the bound in Equation 15.35. This leaves considerably more breathing room. If we take $s = 16$ in place of $s = 15 + 25/512$ we still get a bound which exceeds $1/13$. Thus, this alternate argument completes the proof of the Symmetrization Lemma for all $s \in [12, 16]$.

I imagine that one could continue improving the Symmetrization Lemma by tracking various bounds through a changing domain. For instance, Con-

sider the affine cubes

$$\begin{bmatrix} 416 & 498 \\ 0 & 16 \\ 348 & \mathbf{512} \\ 0 & \pm 24 \end{bmatrix}, \quad \begin{bmatrix} 416 & 498 \\ 0 & 16 \\ 364 & \mathbf{512} \\ 0 & \pm 64 \end{bmatrix}. \quad (15.39)$$

These cubes cover a larger domain. When we use these cubes in place of the ones in §13.4, all the arguments above go through word for word. Thus, Lemma 12.2.2 holds for configurations in which p_1 and p_3 can go all the way up to the unit circle. (Unfortunately, I don't know how to do the same for p_0 and p_2 .) This fact might be helpful for an analysis of what happens to the minimizer for very large power law potentials. See the discussion in §19.1.

Part V

Endgame

Chapter 16

The Baby Energy Theorem

16.1 The Main Result

The goal of this chapter and the next is to prove Lemma 2.4.1. In this chapter we formulate and prove a version of our Energy Theorem which is adapted to the present situation. The result here is much easier to prove, and just comes down to some elementary calculus.

The Domain: Define

$$\Omega = [13, 15 + 25/512] \times [43/64, 1]^2. \quad (16.1)$$

Every configuration in **SMALL4** has the form (p_0, p_1, p_2, p_3) with

$$-p_2 = p_0 = (x, 0), \quad -p_1 = p_3 = (0, y), \quad (x, y) \in [43/64, 1]^2. \quad (16.2)$$

(These conditions define a somewhat larger set of configurations than the ones in **SMALL4**, but we like the conditions above better because they are simpler and more symmetric.) We name such configurations, at the parameter s , by the triple (s, x, y) . The TBP corresponds to the points $(s, 1, 1/\sqrt{3})$, though this point does not lie in Ω .

Blocks: We will use the hull notation developed in connection with our Energy Theorem. See §7.4. So, if X is a finite set of points, $\langle X \rangle$ is the convex hull of X . We say that a *block* is the set of vertices of a rectangle solid, having the following form:

$$X = I \times Q \subset [0, 16] \times [0, 1]^2, \quad (16.3)$$

where I is the set of two endpoints of a parameter interval and Q is the set of vertices of a square. We call X *relevant* if $X \subset \Omega$. We work with the larger domain just to have nice initial conditions for our divide-and-conquer algorithm.

We define

$$|X|_1 = |I|, \quad |X|_2 = |Q|. \quad (16.4)$$

Here $|I|$ is the length of the interval $\langle I \rangle$ and $|Q|$ is the side length of the square $\langle Q \rangle$.

Good Points: We call a point $(x_0, y_0) \in \mathbf{R}^2$ *good* if it names a configuration such that all distances between pairs of points are at least 1.3. The point $(1, 1/\sqrt{3})$ certainly good because all the corresponding distances are at least $\sqrt{2} = 1.41\dots$. For technical reasons having to do with the rationality of our calculation we will end up choosing a good point extremely close to the one representing the TBP and we will work with it rather than the TBP. For this reason, we state our result more generally in terms of good points rather than in terms of the TBP. We fix some unspecified good point $(x_0, y_0) \in \mathbf{R}^2$ throughout the chapter, and in the next chapter we will make our choice.

The Main Result: Let $R_s(x, y)$ denote the Riesz s -energy of the configuration (x, y) . We set $R(s, x, y) = R_s(x, y)$. Given a point $(s, x, y) \in \Omega$, we define

$$\Theta(s, x, y) = R(s, x, y) - R(s, x_0, y_0). \quad (16.5)$$

Our main result is meant to hold with respect to any good point.

Theorem 16.1.1 (Baby Energy) *The following is true for any relevant block X and any good point:*

$$\min_{\langle X \rangle} \Theta \geq \min_X \Theta - (|X|_1^2/512 + |X|_2^2).$$

Remark: It is worth noting the lopsided nature of the Baby Energy Theorem. Thanks to the high exponent in the power laws, our function is much more linear in the s -direction. Our divide-and-conquer operation will exploit this situation.

16.2 A General Estimate

Here we establish well-known facts about smooth functions and their second derivatives.

Lemma 16.2.1 *Suppose $F : [0, 1] \rightarrow \mathbf{R}$ is a smooth function. Then*

$$\min_{[0,1]} F \geq \min(F(0), F(1)) - \frac{1}{8} \max_{[0,1]} |F''(x)|.$$

Proof: Replacing F by $C_1(F - C_2)$ for suitable choices of constants C_1 and C_2 , we can assume without loss of generality that $\min(F(0), F(1)) = 0$ and $|F''| \leq 1$ on $[0, 1]$. We want to show that $F(x) \geq -1/8$ for all $x \in [0, 1]$.

Let $a \in [0, 1]$ be some global minimizer for F . Replacing $F(x)$ by $F(1-x)$ if necessary, we can assume without loss of generality that $a \in [1/2, 1]$. We have $F'(a) = 0$. By Taylor's Theorem with remainder, there is some $b \in [a, 1]$ such that

$$0 \leq F(1) = F(a) + \frac{1}{2} F''(b)(1-a)^2$$

Since $(1-a)^2 \leq 1/4$ the last term on the right lies in $[-1/8, 1/8]$. This forces $F(a) \geq -1/8$. ♠

Lemma 16.2.2 *Let $X \subset \mathbf{R}^n$ denote the set of vertices of a rectangular solid. Let d_1, \dots, d_n denote the side lengths of $\langle X \rangle$. Let $F : \langle X \rangle \rightarrow \mathbf{R}$ be a smooth function. We have*

$$\min_{\langle X \rangle} F \geq \min_X F - \frac{1}{8} \sum_{i=1}^n \left(\max_{\langle X \rangle} |\partial^2 F / \partial x_i^2| \right) d_i^2.$$

Proof: The truth of the result does not change if we translate the domain and/or range, and/or scale the coordinates separately. Thus, it suffices to prove the result when $X = \{0, 1\}^n$, the set of vertices of the unit cube, and $F(0, \dots, 0) = 0$ and $\min_X F = 0$.

We will prove the result by induction on n . The warm-up lemma takes care of the case $n = 1$. Now we consider the n -dimensional case. Consider some point (x_1, \dots, x_n) . We can apply the $(n-1)$ -dimensional result to the

two points $(0, x_2, \dots, x_n)$ and $(1, x_2, \dots, x_n)$. Remembering that $d_i = 1$ for all i , we get

$$F(k, x_2, \dots, x_n) \geq -\frac{1}{8} \sum_{i=2}^n \max_{\langle X \rangle} |\partial^2 F / \partial x_i^2|$$

for $k = 0, 1$. At the same time, we can apply the 1-dimensional result to the line segment connecting our two points. This gives

$$F(x_1, \dots, x_n) \geq \min(F(0, x_2, \dots, x_n), F(1, x_2, \dots, x_n)) - \frac{1}{8} \max_{\langle X \rangle} |\partial^2 F / \partial x_1^2|.$$

Our result now follows from the triangle inequality. ♠

In view of Lemma 16.2.2, the Baby Energy Theorem follows from these two estimates.

1. $|\Theta_{ss}(s, x, y)| \leq 1/64$ for all $(s, x, y) \in \Omega$.
2. $|\Theta_{xx}(s, x, y)| \leq 4$ and $|\Theta_{yy}(s, x, y)| \leq 4$ for all $(s, x, y) \in \Omega$.

Notation: Here we are using the shorthand notation

$$\Theta_{ss} = \frac{\partial^2 \Theta}{\partial s^2}, \quad \Theta_{xx} = \frac{\partial^2 \Theta}{\partial x^2}, \quad \Theta_{yy} = \frac{\partial^2 \Theta}{\partial y^2}. \quad (16.6)$$

In general, we will denote partial derivatives this way. Note that the partial derivative $R_s = \partial R / \partial s$ could be confused with the Riesz energy function at the parameter s . Fortunately, we will not need to consider $\partial R / \partial s$ directly and in the one section where we come close to it we will change our notation.

16.3 A Formula for the Energy

A straightforward calculation shows that

$$R(s, x, y) = A(s, x) + A(s, y) + 2B(s, x) + 2B(s, y) + 4C(s, x, y), \quad (16.7)$$

$$\begin{aligned} A(s, x) &= \left(\frac{(1+x^2)^2}{16x^2} \right)^{s/2} \\ B(s, x) &= \left(\frac{1+x^2}{4} \right)^{s/2} \\ C(s, x) &= \left(\frac{(1+x^2)(1+y^2)}{4(x^2+y^2)} \right)^{s/2} \end{aligned} \quad (16.8)$$

16.4 The First Estimate

Lemma 16.4.1 *Let $\psi(s, b) = b^{-s}$. When $b \geq 1.3$ and $s \in [13, \infty)$ we have $\psi_{ss}(s, b) \in (0, 1/440)$.*

Proof: We compute

$$\psi_{ss}(s, b) = b^{-s} \log(b)^2 > 0. \quad (16.9)$$

Fixing $b \geq 1.3$, this positive quantity is monotone decreasing in s . So, it suffices to prove our result when $s = 13$.

The equation

$$\psi_{ssb}(13, b) = 0$$

has its unique solution in $[1, \infty)$ at $b = \exp(k/13) < 1.3$. Moreover, the function $\psi_{ss}(13, b)$ tends to 0 as $b \rightarrow \infty$. Hence the restriction of $\psi_{ss}(13, *)$ to $[1.3, \infty)$ takes on its max at $b = 1.3$. We compute $|\psi_s(13, 1.3)| < 1/440$. ♠

Let d denote the smallest distance between a pair of points in a configuration parametrized by a point in $\Omega \cup \{x_0, y_0\}$. Finding d is equivalent to finding the global maximum M of the functions A, B, C and then taking $d = M^{-1/2}$. We will see in the next section that we have $M < .59$. The global max in $[43/64, 1]^2$ occurs at the vertex $(43/64, 43/64)$. Since $M < .59$, we have $M^{-1/2} > 1.3$ on Ω . Also, by definition, $M^{-1/2}(x_0, y_0) \geq 1.3$. Hence $d \geq 1.3$. But then, by Lemma 16.4.1, $\Theta_{ss}(s, x, y)$ is the sum of 20 terms, all less than $1/440$ in absolute value.

Now, 10 of the 20 terms comprising $\Theta_{ss}(s, x, y)$ are positive and 10 are negative. Note that 4 of the configuration distances associated to (x, y) are greater or equal to 4 of the configuration distances associated to (x_0, y_0) and *vice versa*. Hence 4 of the positive terms are less or equal to the absolute values of 4 of the negative terms, and *vice versa*. Hence, $|\Theta_{xx}|$ is at most the maximum of the following two quantities..

- The sum of the 6 largest positive terms.
- The sum of the absolute values of the 6 largest negative terms.

In short,

$$|\Theta_{xx}| \leq \frac{6}{440} < \frac{1}{64}.$$

This establishes the estimate.

16.5 A Table of Values

We define

$$a(x) = A(2, x), \quad b(x) = B(2, x), \quad b(x, y) = C(2, x, y). \quad (16.10)$$

These are all rational functions.

We call a function *f* *easy* if the two partial derivatives f_x and f_y are nonzero for all $(x, y) \in [43/64, 1]^2$. For an easy function, the max and min values are achieved at the vertices. A routine exercise in calculus shows that $a, a_x, a_{xx}, b, b_x, b_{xx}, c, c_x$ are all easy. The function c_{xx} is not easy because

$$c_{xxx} = \frac{6x(x^2 - y^2)(-1 + y^4)}{(x^2 + y^2)^4}$$

vanishes along the diagonal. However c_{xxy} does not vanish on our domain and c_{xxx} only vanishes on the diagonal. So, the max and min values of c_{xx} are also achieved at the vertices. Computing at the vertices and rounding outward, we have

$$\begin{array}{lll} a \in [.25, .3] & a_x \in [-.33, 0] & a_{xx} \in [.5, 1.97] \\ b \in [.35, .5] & b_x \in [.34, .5] & b_{xx} \in [.5, .5] \\ c \in [.5, \mathbf{.59}] & c_x \in [-.33, 0] & c_{xx} = [0, .49] \end{array}$$

The value in bold is the one used in the previous section.

Let F be any of A, B, C . We write

$$F = f^u, \quad u = \frac{s}{2}. \quad (16.11)$$

For instance $A = a^u$.

By the chain rule,

$$F_{xx} = u(u-1)f^{u-2}f_x^2 + uf^{u-1}f_{xx} \quad (16.12)$$

This equation will let us get estimates on A_{xx}, B_{xx}, C_{xx} based on the lookup table above. The following technical lemma helps with this goal.

Lemma 16.5.1 *Let $g(u) = ud^{u-1}$ and $h(u) = u(u-1)d^{u-2}$. If $d < .7$ the functions g and h are monotone decreasing as functions of u on $[6.5, \infty)$.*

Proof: Since $g = d/(u - 1) \times h$, it suffices to prove our result for h . Since $h > 0$ the function $\log h$ is well defined. Since \log is a monotone increasing function on \mathbf{R}_+ , it suffices to prove that $\log h(u)$ is decreasing. We compute

$$\frac{d}{du} \log h(u) = \frac{1}{u} + \frac{1}{u-1} + \log(d).$$

Since \log is monotone increasing and $d \leq .7$, we have

$$\frac{d}{du} \log h(u) \leq \frac{1}{6.5} + \frac{1}{5.5} + \log(.7) < -.02.$$

Hence $\log h$ is decreasing on $[6.5, \infty)$ as long as $d \leq .7$. ♠

16.6 The Second Estimate

Lemma 16.6.1 *Through Ω we have*

1. $0 < A_{xx} < 0.035$
2. $0 < B_{xx} < 0.47$
3. $0 < C_{xx} < 0.55$.

Proof: From our lookup table, we see that $f_{xx} \geq 0$ for all $F \in \{A, B, C\}$ and so all of A_{xx}, B_{xx}, C_{xx} are non-negative. Plugging in the values from our lookup table, and using Lemma 16.5.1, we have

$$A_{xx} \leq (13/2)(11/2)(.3)^{9/2}(.11) + (13/2)(.3)^{11/2}(1.97) < 0.035 \quad (16.13)$$

$$B_{xx} \leq (13/2)(11/2)(.5)^{u-2}(.25) + (13/2)(.5)^{11/2}(.5) < 0.47 \quad (16.14)$$

$$C_{xx} \leq (13/2)(11/2)(.59)^{9/2}(.11) + (13/2)(.59)^{11/2}(.49) < 0.55 \quad (16.15)$$

This completes the proof. ♠

Note that $\Theta_{xx} = R_{xx}$ because the good point (x_0, y_0) is not a function of the variable x . By Equation 16.7 we have

$$R_{xx} = A_{xx} + 2B_{xx} + 4C_{xx} \leq .035 + 2(.467) + 4(.55) < 3.2. \quad (16.16)$$

We also have $R_{xx} > 0$. So, $R_{xx} \in (0, 3.2)$, which certainly implies our less precise estimate $|R_{xx}| < 4$. This establishes Estimate 2. The bound for R_{yy} follows from the bound for R_{xx} and symmetry.

The proof of the Baby Energy Theorem is done.

Chapter 17

Divide and Conquer Again

17.1 The Goal

In this chapter we prove Lemma 2.4.1. The first order of business is to make a choice of a good point. Let

$$(x_0, y_0) = \left(1, \frac{37837}{65536}\right). \quad (17.1)$$

We define the function Θ with respect to this point. Note that $65536 = 2^{16}$.

The configuration (x_0, y_0) is extremely close to the TBP on account of the fact that

$$\left| \frac{37837}{65536} - \frac{1}{\sqrt{3}} \right| < 2^{-18}.$$

No configuration isometric to the one represented by Y lies in **SMALL4**. Hence, by the Dimension Reduction Lemma,

$$R_s(Y) > R_s(TBP), \quad \forall s \in \left[13, 15 + \frac{25}{512}\right]. \quad (17.2)$$

We work with (x_0, y_0) rather than $(1, 1/\sqrt{3})$ because we want an entirely rational calculation.

Recall that the set **TINY4** consists of those points $(x, y) \in [55/64, 56/64]$. Lemma 2.4.1 is immediately implied by the following two statements:

1. $\Theta(s, x, y) > 0$ if $(s, x, y) \in \Omega$ and $s \leq 15 + \frac{24}{512}$.
2. $\Theta(s, x, y) > 0$ if $(s, x, y) \in \Omega$ and $s \geq 15 + \frac{24}{512}$ and $(x, y) \in \mathbf{TINY4}$.

17.2 The Ideal Calculation

We first describe our calculation as if it were being done on an ideal computer, without roundoff error.

Given a square Q , we let the vertices of Q be Q_{ij} for $i, j \in \{0, 1\}$. We choose Q_{00} to be the bottom left vertex – i.e., the vertex with the smallest coordinates. Given an interval I we write $I = [s_0, s_1]$, with $s_0 < s_1$.

The Grading Step: Given a block $X = I \times Q \subset [0, 16] \times [0, 1]^2$ we perform the following evaluation,

1. If $I \subset [0, 13]$ we pass X because X is irrelevant.
2. If $I \subset [15 + 25/512, 16]$ we pass X because X is irrelevant.
3. If Q is disjoint from Ω we pass X because X is irrelevant. We test this by checking that either $Q_{10} \leq 43/64$ or $Q_{01} \leq 43/64$.
4. If $s_0 \geq 15 + 24/512$ and $Q \subset \mathbf{TINY4}$ we pass X .
5. $s_0 < 13$ and $s_1 > 13$ we fail X because we don't want to make any computations which involve exponents less than 13.
6. If X has not been passed or failed, we compute

$$\min_X \Theta - |X|_1^2/512 - |X|_2^2.$$

If this quantity is positive we pass X . Otherwise we fail X .

To establish the two statements mentioned at the end of the last section, it suffices to find a partition of $[0, 16] \times [0, 1]^2$ into blocks, all of which pass the grading step.

Subdivision: We will either divide a block X into two pieces or 4, depending on the dimensions. We call X *fat* if

$$16|X|_2 > |X|_1, \tag{17.3}$$

and otherwise thin. If X is fat we subdivide X into 4 equal equal pieces by dyadically subdividing Q . If X is thin we subdivide X into 2 pieces by dyadically subdividing I . We found by trial and error that this scheme takes

advantage of the lopsided form of the Baby Energy Theorem and produces a small partition.

The Main Algorithm: We run the same algorithm as described in §10.2, except for the following differences.

- We start with the initial block $\{0, 16\} \times \{0, 1\}^2$.
- Rather than use some kind of subdivision recommendation coming from the Energy Theorem, we use the fat/thin method just described.

17.3 Floating Point Calculations

When I run the algorithm on my 2016 macbook pro, it finishes in about 0.1 seconds, takes 23149 steps, and produces a partition of size 15481. Figure 17.1 shows a slice of this partition at a parameter $s = 13 + \epsilon$. Here ϵ is a tiny positive number we don't care about. We are showing a region that is just slightly larger than $[1/2, 1]^2$. This is where all the detail is. The grey blocks are irrelevant and the white ones have passed the calculation from the Baby Energy Theorem.

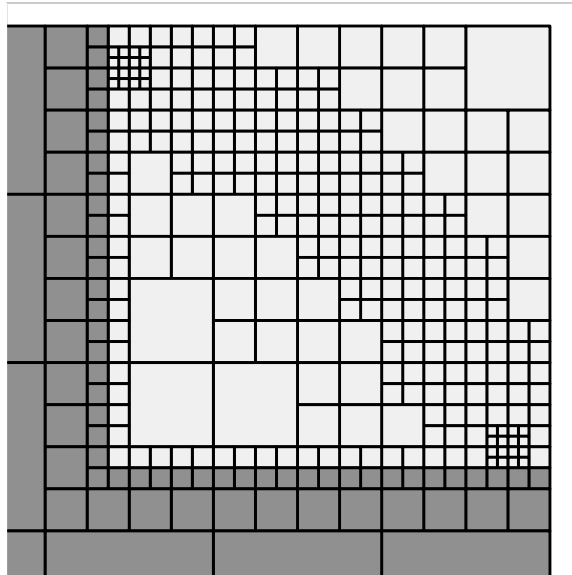


Figure 17.1: Slice of the partition at $s = 13 + \epsilon$.

Figure 17.2 shows a slice of this partition at $s = 15 + 24/512 + \epsilon$. The little squares concentrate around the FPs which are very close to the TBP in terms of energy.

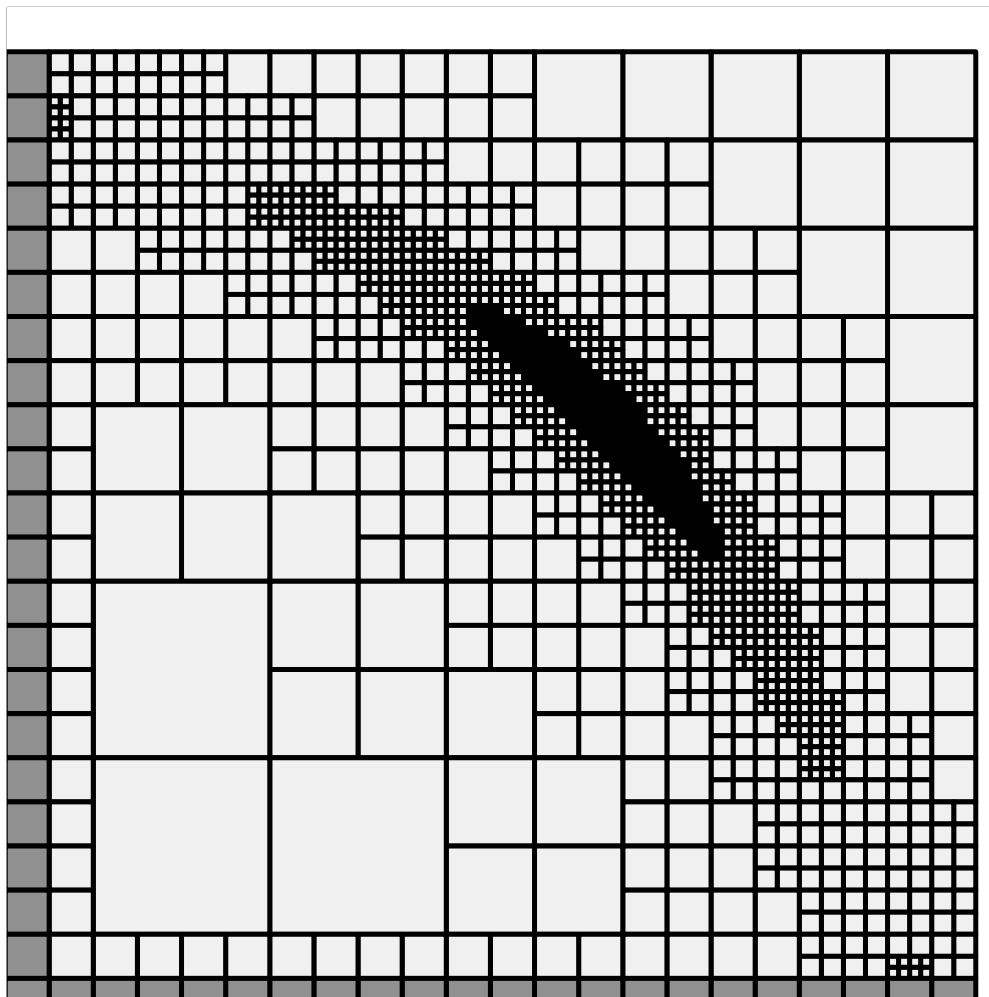


Figure 17.2: Slice of the partition at $s = 13 + \epsilon$.

These pictures don't do the partition justice. Using my program, the reader can slice the partition at any parameter and zoom in and out of the picture.

17.4 Confronting the Power Function

The floating point calculation just discussed does not prove anything, on account of floating point error. The problem is that we need to deal with expressions of the form a^s where a and s are rationals. I don't know how well the floating point errors are regulated for the power function. Without this knowledge, it is difficult to set up an interval arithmetic calculation. Instead, I will explain how to do the calculation using exact integer arithmetic.

First of all, we represent our blocks as integers in the following way:

$$2^{16}X = [S_0, S_1] \times [X_0, X_1] \times [Y_0, Y_1], \quad S_i, X_j, Y_k \in \mathbf{Z}. \quad (17.4)$$

In theory, the subdivision process could render this representation invalid. This would happen, say, if $S_0 + S_1$ is odd and we have to subdivide X along the interval $I = [S_0/2^{16}, S_1/2^{16}]$. However, our algorithm comes nowhere near doing enough nested subdivisions for this to happen.

Referring to the functions in §16.3, in order to evaluate the functions $F_k(s, x, y)$ for $s, x, y \in \mathbf{Q}$ we need to be able to handle expressions of the form

$$b^s, \quad b, s \in \mathbf{Q}. \quad (17.5)$$

Now we explain how we do this.

A Specialized Power Function: We introduce the function

$$P(\sigma, b, s), \quad \sigma \in \{-1, 1\}. \quad (17.6)$$

Here is the definition:

1. We compute the floating point value $y = b^s$.
2. We compute the floating point value $2^{32}y$ and cast it to an integer N . The casting process rounds this value to a nearby integer – presumably one of the two nearest choices.
3. $P(\sigma, b, s) = 2^{-32}(N + 2\sigma)$. (Here 2σ is deliberate.)

It seems possible that the value $P(\sigma, b, s)$ might be computer-dependent, because of the way that $2^{32}y$ is rounded to an integer.

Heuristic Discussion: Though this is irrelevant for the formal proof, let me first justify heuristically why one should expect

$$P(-1, b, s) < b^s < (+1, b, s), \quad (17.7)$$

on any modern computer, at least for smallish rationals b and s .

Unless we had an unbelievably poor computer, we should expect

$$|N - 2^{32}y| < 1.5,$$

because the computer is supposed to choose an integer within 1 unit of $2^{32}y$. At the same time, unless we had an extremely poor computer, we should expect

$$|y - b^s| < .5 \times 2^{-32},$$

because otherwise the power function would be off in roughly the 10th digit.

Assuming these equations, we have

$$(N - 2) - 2^{32}y < -.5,$$

which implies

$$(N - 2) - 2^{32}b^s < (N - 2) - 2^{32}y + .5 < 0.$$

But this gives $P(-1, b, s) < b^s$. Similarly, we should expect $P(1, b, s) > b^s$.

Formal Verification: The above heuristics do not play any role in our proof, but we mentioned them in order to explain the idea behind the construction. What we actually do is compute $P(\pm 1, b, s)$ and then formally verify Equation 17.7 for the relevant numbers. Here we explain how we check that either $p < b^s$ or $b^s < p$ for rational numbers p, b, s using integer arithmetic.

We define integers $\alpha, \beta, \gamma, \delta, u, v$ such that

$$p = \frac{\alpha}{\beta}, \quad b = \frac{\gamma}{\delta}, \quad s = \frac{u}{v}. \quad (17.8)$$

The sign of $p - b^s$ is the same as the sign of

$$\alpha^v \delta^u - \beta^v \gamma^u. \quad (17.9)$$

We just expand out these integers and check the sign. We implement the verification using Java's BigInteger class, which performs exact integer arithmetic on integers whose size is limited only by the capacity of the machine. In theory, the needed verification could be larger than the capacity of the machine, but for our algorithm, the calculation runs to completion, doing all verifications in less than about 2 minutes.

17.5 The Rational Calculation

Now we describe the calculation we actually do. We represent our blocks as in the preceding section.

For each of the functions F_k in §16.3 we define \underline{F}_k and \overline{F}_k to be the functions obtained by replacing the operation b^s with $P(-1, b, s)$ and $P(1, b, s)$ respectively. For instance

$$\underline{F}_2(s, x) = P\left(-1, \frac{1+x^2}{4}, s/2\right). \quad (17.10)$$

We then define \underline{R} and \overline{R} as in Equation 16.7 except that we replace each occurrence of a function F with \underline{F} or \overline{F} . For each instance we actually compute, we formally verify all relevant instances of Equation 17.7. (We set the calculation to abort with a conspicuous notification if any of these formal verifications fails.) These verifications give us the inequalities

$$\underline{R}(s, x, y) < R(s, x, y) < \overline{R}(s, x, y) \quad (17.11)$$

in every instance that we need it.

We define

$$\underline{\Theta}(s, x, y) = \underline{R}(s, x, y) - \overline{R}(s, x_0, y_0). \quad (17.12)$$

In all the cases we need, we have

$$\underline{\Theta}(s, x, y) < \Theta(s, x, y). \quad (17.13)$$

Finally, in Step 5 of the Grading Algorithm, we pass the block X if we have

$$\min_X \underline{\Theta} - |X|_1^2/512 - |X|_2^2 > 0. \quad (17.14)$$

For all the values we need, this equation guarantees that

$$\min_X \Theta - |X|_1^2/512 - |X|_2^2 > 0,$$

which combines with the Baby Energy Theorem to show that $\Theta > 0$ on $\langle X \rangle$.

When we re-run our calculation using this rational approach, the calculation runs to completion in about 2 minutes and 15 seconds. This is about 1500 times as long as the floating point calculation. The rational calculation takes 23213 steps and produces a partition of size 15519. This partition is slightly larger than the one produced by the floating point calculation because the rational version of Grading Step 6 is slightly more stringent.

The fact that the rational version of our program runs to completion constitutes a proof of the two statements mentioned at the end of §17.1, and thereby furnishes a proof of Lemma 2.4.1.

17.6 Discussion

It seems worth mentioning some other approaches one could take when dealing with the calculations described in this chapter. First of all, the algorithm halts in about 23000 steps, and perhaps there are about a half a million times we must invoke the power function. This is a pretty small number in modern computing terms, and so we could have simply made a lookup table, where we store the needed values of the power function in a file and let the computation access the file as needed. This approach is somewhat like the construction of our function $P(\sigma, b, s)$, except that the values would be frozen once and for all, and someone could go back and verify that the values stored in the file had the claimed properties by any means available.

Another approach would be to observe that we are only considering expressions of the form b^s where $b \in Q$ and $s = a/2^k$. That is, s is a dyadic rational. We could consider the identity

$$b^s = \sqrt{\dots \sqrt{b \times \dots \times b}},$$

which only involves multiplication and the square root function, both of which are governed by the IEEE standards. One could use this as a basis for an interval arithmetic calculation, though I don't know how efficient it would be. Like our formal verifier, the success of this approach would depend on k being fairly small.

Yet another approach would be to construct polynomial over and under approximations to the function $p(b, s) = b^s$, somewhat along the lines of §4. This would be a more robust approach, because it wouldn't depend on the denominator of s being fairly small.

Chapter 18

The Tiny Piece

18.1 Overview

Define

$$\Omega' = \left[15 + \frac{24}{512}, 15 + \frac{25}{512}\right] \times \left[\frac{55}{64}, \frac{56}{64}\right]. \quad (18.1)$$

Lemma 2.4.2 concerns triples $(s, x, y) \in \Omega'$. As in the previous chapters, we have our Riesz energy function $R : \Omega' \rightarrow \mathbf{R}$, given by $R(s, x, y) = R_s(x, y)$.

We consider the curves

$$\gamma(x_0, t) = (x_0 + t - 2t^2, x_0 - t - 2t^2). \quad (18.2)$$

Lemma 2.4.2 says that

$$R(s, \gamma(x_0, 0)) \leq R(s, \gamma(x_0, t)), \quad \forall (x, \gamma(x_0, t)) \in \Omega', \quad (18.3)$$

with equality iff $t = 0$.

We've mentioned already that Equation 18.3 is delicate. The result does not work, for instance, if we forget the quadratic term and just use the more straightforward retraction along lines of slope -1 in **TINY4**. The trouble is that the expression

$$R_{xx} + R_{yy} - 2R_{xy} \quad (18.4)$$

is negative at some points of Ω' . The straight line retraction would actually work if we could use a region much smaller than Ω' but I didn't think I could prove a much stronger version of Lemma 2.4.1. The more complicated retraction has nicer properties which allow it to work throughout Ω' .

18.2 The Proof Modulo Derivative Bounds

When we twice differentiate Equation 18.3 with respect to t we get an expression that is more complicated than Equation 18.4, and the other terms turn out to be helpful. What we will prove is that $R_{tt} > 0$ throughout Ω' . This means that the function $t \rightarrow R(s, \gamma(x_0, t))$ is convex as long as the image lies in Ω' . Since $R(s, \gamma(x_0, t)) = R(s, \gamma(x_0, -t))$, it follows from symmetry and convexity that R takes its minimum on such curves when $t = 0$. In short, Lemma 2.4.2 follows directly from the fact that $R_{tt} > 0$ on Ω' .

Using the Chain Rule (or, more honestly, Mathematica) we compute that

$$R_{tt} = (1 - 8t + 16t^2)R_{xx} + (1 + 8t + 16t^2)R_{yy} + (-2 + 32t^2)R_{xy} - 4(R_x + R_y).$$

There is nothing special about the function R in this equation. We would get the same result for any map $t \rightarrow F(s, \gamma(x_0, t))$.

Now we simplify the expression that we need to consider. It follows from Equation 16.7 and the results in Lemma 16.6.1 that $R_{xx} > 0$ and $R_{yy} > 0$ in Ω' . We will prove below in Lemma 18.3.1 that $R_{xy} > 0$ on Ω' as well.

Since $R_{xx} > 0$ and $R_{yy} > 0$ and $R_{xy} > 0$ on Ω' , we see that

$$R_{tt} \geq (1 - 8t)R_{xx} + (1 + 8t)R_{yy} - 2R_{xy} - 4(R_x + R_y).$$

Since $2t = x - y$, see that $R_{tt} \geq H$, where

$$H = (1 - 4x + 4y)R_{xx} + (1 + 4x - 4y)R_{yy} - 2R_{xy} - 4(R_x + R_y). \quad (18.5)$$

We will complete the proof of Lemma 2.4.2 by showing that $H > 0$ on Ω' .

Since (as it turns out) H is bounded away from 0 on Ω' , we can get away with just using information about the first partial derivatives. The goal of the next three sections is to prove the following estimates.

$$|H_x|, |H_y| < 16, \quad H_s < 2. \quad (18.6)$$

These are meant to hold in Ω' . We are not ruling out the possibility that H_s could be very negative, though a further examination of our argument would rule this out. Also, for what it is worth, a careful examination of the proofs will show that really we prove these bounds on the larger domain $[15, 16] \times \mathbf{TINY4}$.

We set

$$\beta = 15 + \frac{25}{512}. \quad (18.7)$$

The bound on H_s tells us that

$$H(x, y, s) > H(\beta, x, y) - \frac{1}{256}, \quad \forall (s, x, y) \in \Omega'. \quad (18.8)$$

So, to finish the proof of Lemma 2.4.2 we just have to show that

$$H(\beta, x, y) > \frac{1}{256}, \quad \forall x, y \in [55/64, 56/64]. \quad (18.9)$$

This reduces our calculation down to a single parameter.

We compute, for the 17^2 points

$$x^*, y^* \in \{55/64 + i/1024, i = 0, \dots, 16\} \quad (18.10)$$

that

$$H(\beta, x^*, y^*) > \frac{1}{64} + \frac{1}{256}. \quad (18.11)$$

We do this calculation in Java using exact integer arithmetic, exactly as we did the calculation for Lemma 2.4.1.

Given any point (β, x, y) there is some point (x^*, y^*) on our list such that $|x - x^*| \leq 2048$ and $|y - y^*| \leq 2048$. But then the bounds $|H_x| < 16$ and $|H_y| < 16$ give us

$$H(\beta, x, y) > H(\beta, x, y^*) - \frac{1}{128} > H(\beta, x^*, y^*) - \frac{1}{64} > \frac{1}{256}.$$

This establishes Equation 18.9. Our proof of Lemma 2.4.2 is done, modulo the derivative bounds in Equation 18.6. The rest of the chapter is devoted to establishing these bounds.

The bound on $|H_x|$ implies the bound on $|H_y|$ by symmetry. So, we just have to deal with $|H_x|$ and $|H_s|$. Before we launch into the tedious calculations, we note that a precise estimate like $|H_x| < 16$ is not so important for our overall proof. Were we to have the weaker estimates $|H_x|, |H_y| < 32$ we just need to compute at 4 times as many values. Given the speed of the computation above (less than 30 seconds) we would have a feasible calculation with much worse estimates. We mention this because, even though we tried very hard to avoid any errors of arithmetic, we don't want to the reader to think that the proof is so fragile that a tiny error in arithmetic would destroy it.

18.3 The First Bound

We first prepare another table like the one in §16.5. This time we compute on the smaller domain

$$\mathbf{TINY4} = [55/64, 56/64].$$

We define the functions a, b, c as in §16.5. In addition to the functions considered in §16.5, the functions $a_{xxx}, b_{xxx}, c_{xy}, c_{xxx}, c_{xxy}$ are also easy on **TINY4**. We have

$$\begin{array}{llll} a \in [.254, .256] & a_x \in [-.09, -.07] & a_{xx} \in [.76, .83] & a_{xxx} \in [-3.3, -2.9] \\ b \in [.43, .45] & b_x \in [.42, .44] & b_{xx} \in [.5, .5] & b_{xxx} \in [0, 0]. \\ c \in [.50, .52] & c_x \in [-.09, -.07] & c_{xx} \in [.08, .11] & c_{xxx} \in [-.012, .012] \\ c_{xy} \in [.67, .71] & c_{xxy} \in [-1.07, -.97] & & \end{array}$$

Note that $b_{xx} = 1/2$ identically and $b_{xxx} = 0$ identically. One can deduce other bounds by symmetry. For instance, the bounds on c_y are the same as the bounds on c_x . Now we use these bounds to get derivative estimates on the functions A, B, C and R . We already mentioned the inequality $0 < R_{xy}$ above. This is part of our next result.

Lemma 18.3.1 $0 < R_{xy} < .35$ in Ω' .

Proof: Since the functions A and B in Equation 16.7 only involve one of the variables, we have $R_{xy} = 4C_{xy}$. We will show that $0 < C_{xy} < .087$ in Ω' . As in Lemma 16.6.1 we write $u = s/2$ and $C = c^u$.

We compute

$$C_{xy} = u(u-1)c^{u-2}c_xc_y + uc^{u-1}c_{xy} \quad (18.12)$$

Since $c_x < 0$ and $c_y < 0$ and $c_{xy} > 0$, we have $C_{xy} > 0$. Given the bounds in our table, and Lemma 16.5.1, we have

$$C_{xy} < (15/2) \times (14/2) \times (.52)^{11/2} \times (.09)^2 + (15/2) \times (.52)^{13/2} (.71) < .087.$$

This completes the proof. ♠

Now we revisit Lemma 16.6.1, using the values from the smaller domain to get better estimates.

Lemma 18.3.2 *Throught Ω' we have*

1. $0 < A_{xx} < 0.002$
2. $0 < B_{xx} < 0.138$
3. $0 < C_{xx} < 0.023$.

Proof: The positivity follows from Lemma 16.6.1. For the new upper bounds, we proceed just as in Lemma 16.6.1 except that now we have different values to plug in. Using Lemma 16.5.1 and our lookup table, we have

$$A_{xx} \leq (15/2)(13/2)(.256)^{11/2}(.09)^2(15/2)(.256)^{13/2}(.83) < .002 \quad (18.13)$$

$$B_{xx} \leq (15/2)(13/2)(.45)^{11/2}(.44)^2 + 15/2(.45)^{13/2}(.5) < .138 \quad (18.14)$$

$$C_{xx} \leq (15/2)(13/2)(.52)^{11/2}(.09)^2 + 15/2(.45)^{13/2}(.11) < .023 \quad (18.15)$$



Corollary 18.3.3 $|R_{xx}|, |R_{xx}| < .37$ on Ω' .

Proof: From 16.7 we get

$$|R_{xx}| \leq |A_{xx}| + 2|B_{xx}| + 4|C_{xx}| \leq .002 + 2(.138) + 4(.023) < .37. \quad (18.16)$$



Lemma 18.3.4 *The following is true in Ω' .*

1. $|A_{xxx}| < .011$.
2. $|B_{xxx}| < .77$.
3. $|C_{xxx}| < .052$.

$$4. |C_{xy}| < .31.$$

Proof: As usual we set $u = s/2$ and $F = f^u$ for each $F = A, B, C$. We have

$$F_{xxx} = u(u-1)(u-2)f^{u-3}f_x^3 + 3u(u-1)f^{u-2}f_xf_{xx} + uf^{u-1}f_{xxx}.$$

Hence

$$\begin{aligned} |F_{xxx}| &\leq \\ &u(u-1)(u-2)f^{u-3}|f_x|^3 + \\ &3u(u-1)f^{u-2}|f_x||f_{xx}| + \\ &uf^{u-1}|f_{xxx}|. \end{aligned} \tag{18.17}$$

Plugging in the max values from our charts, we get the bounds in Items 1-3.

A similar calculation gives

$$\begin{aligned} |F_{xy}| &\leq \\ &u(u-1)(u-2)f^{u-3}|f_x|^2|f_y| + \\ &2u(u-1)f^{u-2}|f_x||f_{xy}| + \\ &u(u-1)f^{u-2}|f_y||f_{xx}| + \\ &uf^{u-1}|f_{xxx}|. \end{aligned} \tag{18.18}$$

Plugging in the max values from our charts, we get the bound in Item 4. ♠

Corollary 18.3.5 $|R_{xxx}|, |R_{yyy}| < 3.3$ and $|R_{xy}|, |R_{yx}| < 1.24$ on Ω' .

Proof: Using the bounds in the preceding lemma, we have

$$|R_{xxx}| \leq 2|A_{xxx}| + 4|B_{xxx}| + 4|C_{xxx}| < 2(.011) + 4(.76) + 4(.052) < 3.3$$

Similarly

$$|R_{xy}| = 4|C_{xy}| < 1.24.$$

The other two cases follow from symmetry. ♠

Now for the bound on $|H_x|$. We first remind the reader of the equation for H .

$$H = (1 - 4x + 4y)R_{xx} + (1 + 4x - 4y)R_{yy} - 2R_{xy} - 4(R_x + R_y). \quad (18.19)$$

We compute

$$H_x = (1 + 4x - 4y)R_{xxx} - 4R_{xx} + (1 - 4x + 4y)R_{xyy} + 4R_{yy} - 2R_{xxy} - 4R_{xx} - 4R_{xy}.$$

Noting that $|4x - 4y| \leq 1/32$ on Ω' , and using the bounds in this section, we get

$$|H_x| < \frac{33}{32}(3.3) + 4(.37) + \frac{33}{32}(1.24) + 4(.37) + 2(1.24) + 4(.37) + 4(.35) < 16.$$

This completes the proof.

18.4 The Second Bound

For most of the calculations we use the parameter $u = s/2$ in our calculations. Note that $F_s = 2F_u$ for any function F . One principle we use repeatedly is that $|\log(x)|$ is monotone decreasing when $x \in (0, 1)$.

Lemma 18.4.1 $A_{xu}, C_{xu} > 0$ and $|B_{xu}| < .015$ on Ω .

Proof: Letting $F = f^u$ be any of our functions, we have $F_x = uf^{u-1}f_x$. Differentiating with respect to u , and noting that $f_{xu} = 0$ because f does not depend on u , we have

$$F_{xu} = f^{u-1}f_x \times (1 + u \log f). \quad (18.20)$$

In all cases we have $\log f \in [-1.35, -.65]$ and this combines $u \in [15/2, 8]$ to force the $1 + u \log f < 0$.

Since $a_x, c_x < 0$ we see that $A_{xu}, C_{xu} > 0$. Since $u \leq 8$, we have

$$|1 + u \log b| < 8|\log(.42)| - 1 < 6.$$

Since $u > 15/2$, Lemma 16.5.1 and the bounds in the lookup table give

$$|B_{xu}| < (6)(.45)^{13/2}(.44) < .015.$$

This completes the proof. ♠

Corollary 18.4.2 $-4(R_{xs} + R_{ys}) < 1$.

Proof: R_{xu} is the sum of 10 terms. 6 of these terms are positive and the other 4 are greater than $-.015$. Hence $R_{xu} > -.06$. By symmetry $R_{yu} > -.06$. Hence

$$-4(R_{xs} + R_{ys}) = -8(R_{xu} + R_{yu}) < 8 \times .12 < 1$$

This completes the proof. ♠

Lemma 18.4.3 $A_{xxu}, B_{xxu}, C_{xxu} < 0$.

Proof: Here is Equation 16.12 again:

$$F_{xx} = u(u-1)f^{u-2}f_x^2 + uf^{u-1}f_{xx}.$$

Differentiating with respect to u , we have

$$F_{xxu} = f^{u-2}f_x^2 \times (2u-1 + u(u-1)\log f) + f^{u-1}f_{xx} \times (1 + u\log f). \quad (18.21)$$

The same analysis as in the preceding lemma shows that

$$(2u-1) + u(u-1)\log f < 0, \quad (1 + u\log f) < 0.$$

Again, what is going on is that $\log f$ is not too far from -1 , and the coefficient in front of $\log f$ is much larger than the constant term.

Hence, in all cases, the first term on the right hand side of Equation 18.21 is negative. Since $f_{xx} > 0$, the second term is also negative. ♠

Corollary 18.4.4 $(1 + 4x - 4y)R_{xxs} + (1 + 4x - 4y)R_{yys} < 0$ on Ω' .

Proof: R_{xxu} is the sum of 10 terms all negative. Hence $R_{xxu} < 0$. By symmetry $R_{yyu} < 0$ as well. But then the same goes for R_{xxs} and R_{yys} . Moreover $|4x - 4y| \leq 1/128$ on Ω' , so the coefficients in front of our negative quantities are positive. ♠

Remark: Of course, we have also proved that $R_{ssx}, R_{yys} < 0$. This fact will be useful in the next chapter.

Lemma 18.4.5 $C_{xyu} \in [-.053, 0)$ on Ω' .

Proof: Treating Equation 18.12 just as we treated Equation 16.12, we have

$$C_{xyu} = c^{u-2}c_xc_y \times (2u - 1 + u(u - 1) \log c) + c^{u-1}c_{xy} \times (1 + u \log c). \quad (18.22)$$

From the bounds in the lookup table, we get $\log(c) \in [-.70, -.65]$. From this fact, we get the bounds

$$(2u - 1 + u(u - 1) \log c) \in (-25.2, 0), \quad (1 - u \log c) \in (-4.6, 0).$$

Since $c_xc_y > 0$ and $c_{xy} > 0$ we see that $C_{xyu} < 0$. Finally, since $u > 15/2$, Lemma 16.5.1 combines with the values in the table to give

$$|C_{xyu}| < (25.2)(.52)^{11/2}(.09)^2 + (4.6)(.52)^{13/2}(.71) < .053.$$

This completes the proof. ♠

Corollary 18.4.6 $-2R_{xys} < 1$ on Ω' .

Proof: We have $|R_{xyu}| = 4|C_{xyu}| < .22$. Hence $|2R_{xys}| = 4 \times .22 < 1$. ♠

Our bound $H_s < 2$ follows from adding the bounds in the 3 corollaries. Our proof of Lemma 2.4.2 is now complete. We mention one more corollary that we will use in the next chapter.

Corollary 18.4.7 $R_{xys} < 0$ on Ω' .

Proof: We have $R_{xys} = 2R_{xyu} = 8C_{xyu} < 0$. ♠

Chapter 19

The Final Battle

19.1 Discussion

Before getting into the details of our proof, it seems worth pausing to give some geometric intuition about the competition between the TBP and the FPs. Suppose we start with the *equatorial* FP, the one which has its apex at the north pole and 4 points at the equator. Near the equator the sphere is well approximated by an osculating cylinder. If we think of the points as living on the cylinder instead of the sphere, then pushing the 4 points down dramatically increases the distance to the apex without changing any of the other distances. When the exponent in the power law is very high, this motion gives a tremendous savings in the energy. When we do the motion on the sphere, the tiny decrease in some of the distances is a much smaller effect relatively speaking. This mismatch between the dramatic energy decrease in some bonds and a small energy increase in others is enough for an FP to overtake the TBP as the energy minimizer. The question is really just when this happens.

One interesting thing to observe is the TBP beats any given FP in the long run. That is, if X is any TBP, we have $R_s(X) > R_s(TBP)$ for all s sufficiently large. In all cases other than the equatorial FP, the minimum distance between a pair of points in X is less than the minimum distance between pairs of points in the TBP. In the long run, these minimum distances dominate in the energy calculation. (For the equatorial FP, a direct calculation bears out that the TBP always has lower energy.) Put another way, the amount we need to push the points downward from the equatorial

FP described above decreases as the power law increases. As $s \rightarrow \infty$, the winning FP converges to the equatorial FP.

Finally, here are some thoughts on how to continue our analysis into the realm of very large exponents. In playing around with the Symmetrization Lemma from Part 3, I noticed that as the power law increases, the domain in which our proof of the Symmetrization Lemma seems to work stretches out to include points near the axes and closer to the unit circle. Thus, the “domain of proof” for the Symmetrization Lemma somehow follows the domain of interest in the minimization problem. See the discussion at the end of §15.5. Perhaps this means that a generalization of the Symmetrization Lemma would prove that an FP is the global minimizer w.r.t. R_s when $s > 15 + 25/512$.

19.2 Proof Overview

Now we turn to the proof of the Main Theorem. Recall that

$$\alpha = 15 + \frac{24}{512}, \quad \beta = 15 + \frac{25}{512}. \quad (19.1)$$

We now know that

- The TBP is the unique minimizer w.r.t R_s when $s \in (0, \alpha]$.
- When $s \in [\alpha, \beta]$. the minimizer w.r.t R_s is either the TBP or an FP. In the latter case, the FP lies in **TINY4**.

It remains only to compare the TBP with the FPs and see what happens. We divide the problem into 4 regimes:

1. The critical interval, $[\alpha, \beta]$. We will prove a monotonicity result which allows us to detect the phase-transition constant $\mathfrak{w} \in (\alpha, \beta)$.
2. The small interval, $[\beta, 16]$. We show that the TBP is never the minimizer w.r.t R_s when s is in this interval. We do this with a 6-term Taylor series expansion of the energy about the point $(\beta, 445/512, 445/512)$.
3. The long interval $[16, 24]$. We do this with a carefully engineered brute-force calculation of the energy at 64 specific points.

4. The end $[24, \infty)$. Here we give an analytic argument that is inspired by the geometric picture discussed above.

In this chapter we will use Mathematica's ability to do very high precision approximations of logs of rational numbers, and rational powers of rational numbers. We will list our calculations with about 40 digits of accuracy, and the reader will see that usually we only need a few digits of accuracy, and in all cases at most 8 digits. The calculations are so concrete that a reader who does not trust Mathematica can re-do the calculations using some other system. One could also use methods such as those in §4 to do the evaluations.

19.3 The Critical Interval

Let Ω' be the region from the previous chapter. Let $\Omega'' \subset \Omega'$ be the rectangle of points (s, x, x) . These points parametrize the FPs in **TINY4**.

Let $\Theta : \Omega' \rightarrow \mathbf{R}$ be the function considered in the proof of Lemma 2.2.2, except that this time we define Θ w.r.t the good point $(x_0, y_0) = (1, 1/\sqrt{3})$ corresponding to the TBP. In other words,

$$\Theta(s, x, x) = R(s, x, x) - R(s, 1, 1/\sqrt{3}). \quad (19.2)$$

We consider the partial derivative.

$$\Psi = \Theta_s. \quad (19.3)$$

Lemma 19.3.1 (Negative Differential) $\Psi < 0$ on Ω' .

Proof: In §16 we proved that

$$|\Psi_s| < 2^{-6}. \quad (19.4)$$

Given this bound, and the fact that every parameter associated to Ω' is within 2^{-9} of s^* , it suffices to prove that

$$\Psi(\alpha, x, x) < -2^{-15} = .0000305..., \quad \forall x \in [55/64, 56/64]. \quad (19.5)$$

We have already shown that $R_{xxs} < 0$ and $R_{yys} < 0$ on Ω' . See the remark following Corollary 18.4.4. Moreover, Lemma 18.4.7 says that $R_{xys} < 0$. But now we observe that

$$\Psi_{xx} = R_{xxs} + R_{yys} + 2R_{xys} < 0. \quad (19.6)$$

In other words, the single variable function

$$\zeta(x) = \Psi(\alpha, x, x) \quad (19.7)$$

is concave. That is, $\zeta''(x) < 0$.

We compute explicitly in Mathematica that

$$\zeta'(55/64) = -0.00134860003628414037588921953585756185...$$

$$\zeta'(56/64) = -0.00625293301806721131316180538457106515...$$

Our point here is that Mathematica can approximate the quantities involved to arbitrarily high precision. All we care about is that these quantities are both negative, and for this we just need 3 digits of accuracy.

Since ζ' is negative at both endpoint and $\zeta'' > 0$ we see that

$$\zeta'(x) < 0, \quad \forall x \in [55/64, 56/64]. \quad (19.8)$$

Hence ζ takes its maximum in our interval at $55/64$. Again using Mathematica, we compute that

$$\psi(55/64) = -0.000112582444773451000945188641412... < -2^{-15}.$$

Since this is the maximum value on our interval, Equation 19.5 is true. ♠

Now we reach the point where we define the phase transition constant ϖ . We don't get an explicit value, of course, but we get the existence from the Negative Differential Lemma.

Lemma 19.3.2 *There exists a constant $\varpi \in (\alpha, \beta)$ such that*

1. *For $s \in (0, \varpi)$ the TBP is the unique global minimizer w.r.t R_s .*
2. *For $s = \varpi$, both the TBP and some FP are minimizers w.r.t R_s .*
3. *For $s \in (\varpi, \beta]$, the minimizer w.r.t R_s is an FP.*

Proof: Combining the Big Theorem, the Small Theorem, the Symmetrization Lemma, and Lemma 2.4.1, we can say that the TBP is the unique global minimizer w.r.t R_s for all $s \in (0, \alpha]$.

On the other hand, we compute in Mathematica that

$$\Theta\left(\beta, \frac{445}{512}, \frac{445}{512}\right) < -.00000079529630852573464048109... \quad (19.9)$$

So, the TBP is not the minimizer w.r.t. R_s when $s = \beta$.

These two facts combine with Lemma 2.4.2 to say that there exists a constant $\mathfrak{w} \in (\alpha, \beta)$ which has Properties 1 and 2. But for $s \in (\mathfrak{w}, \beta]$ we have

$$\Theta\left(s, \frac{445}{512}, \frac{445}{512}\right) = \int_{\mathfrak{w}}^s \frac{d\Theta}{d\mu} d\mu < 0, \quad (19.10)$$

by the Negative Differential Lemma. So, for all such s , some FP beats the TBP. This establishes Property 3. ♠

Remark: Of course, we found the value 445/512 by trial and error. No other dyadic rational with denominator less than 1024 works.

19.4 The Small Interval

Our goal in this section is to prove that the TBP is not a minimizer w.r.t R_s when $s \in [\beta, 16]$. We change our notation for partial derivatives to one which is more convenient for the present purposes. We write

$$\partial_s F = \partial F / \partial s, \quad \partial_s^2 F = \partial^2 F / \partial s^2, \dots \quad (19.11)$$

for partial derivatives, rather than F_s , F_{ss} , etc, as we have been doing. We have $\Psi = \partial\Theta/\partial s$, as above.

We start this section by getting good estimates on Ψ restricted to the ray

$$\left[\beta, \infty\right) \times \{(x^*, x^*)\}, \quad x^* = \frac{445}{512}. \quad (19.12)$$

The idea is to expand things out in a Taylor series about the point (β, x^*, x^*) .

Let $\{b_{ij}\}$ denote the set of 10 distances which arise in connection with the TBP and let $\{b_{ij}^*\}$ denote the set of 10 distances which arise in connection with the FP corresponding to (x^*, x^*) . We have

$$\partial_s^k \Psi(\beta, x^*, x^*) = \sum_{i < j} (-1)^{k+1} (\log b_{ij}^*)^{k+1} (b_{ij}^*)^{-\beta} - \sum_{i < j} (-1)^{k+1} (\log b_{ij})^{k+1} b_{ij}^{-\beta}. \quad (19.13)$$

Note that $b_{ij} > 1$ and $b_{ij}^* > 1$ for all i, j and so the two summands have opposite signs, and the signs alternate with k . Both summands decrease in absolute value with s . Therefore,

$$|\partial_s^k \Psi(s, x^*, x^*)| \leq \max \left(\left| \sum_{i < j} (\log b_{ij}^*)^{k+1} (b_{ij}^*)^{-\beta} \right|, \left| \sum_{i < j} (\log b_{ij})^{k+1} b_{ij}^{-\beta} \right| \right). \quad (19.14)$$

Call this bound M_k .

We set

$$C_k = \frac{1}{k!} \partial_k \Psi(\beta, x_0, y_0), \quad t = s - \beta. \quad (19.15)$$

From Taylor's Theorem with Remainder, we have

$$\Psi(s, x^*, y^*) \leq C_0 + C_1 t + C_2 t^2 + C_3 t^3 + C_4 t^4 + (M_5/5!) t^5, \quad (19.16)$$

Using Mathematica, we compute

$$\begin{aligned} C_0 &= -0.0001406516656256745696597771130279382483753... \\ C_1 &= +0.0001070798482749202926537278987632143748257... \\ C_3 &= -0.0000289210547274380613542311596565183497424... \\ C_4 &= +0.0000042894047365987515852310727996101178313... \\ C_5 &= -0.0000003935379322252784711450416897633466139... \\ M_5/5! &= +0.0000006949816058762526443113985497276969081... \end{aligned}$$

From these values we see that

$$-|C_0|^{-1} \Psi(s) \geq 1 - .77t + .2t^2 - .031t^3 + .003t^4 - .005t^5. \quad (19.17)$$

The polynomial on the right is positive dominant in the sense of §5. Hence, it is positive for $t \in [0, 1]$. Therefore,

$$\frac{\partial \Theta}{\partial s} = \Psi(s, x^*, x^*) < 0, \quad s \in [\beta, \beta + 1]. \quad (19.18)$$

But we also know that $\Theta(\beta, x^*, x^*) < 0$. Hence $\Theta(s, x^*, x^*) < 0$ for all $s \in [\beta, 16]$. Hence the TBP is not a minimizer w.r.t. R_s when $s \in [\beta, 16]$.

19.5 The Long Interval

Now we prove that the TBP is not a minimizer w.r.t. R_s when $s \in [16, 24]$. As we mentioned in the beginning of the chapter, the winning FP keeps

changing. So, we have to adjust our domain. Now we work in the domain $[16, 24] \times [7/8, 1)^2$. Referring to the argument in §16.4, the smallest distance that arises in connection with configurations corresponding to points in this domain is

$$\left(C(7/8, 7/8)\right)^{-1/2} = \frac{112}{113}\sqrt{2} > 7/5.$$

This means that each individual summand in the 20 terms which go into $\partial_s^2 \Theta$ is at most $(7/5)^{-s} \log(7/5)^2$ in absolute value. Reasoning as in §16.4, we have

$$|\partial_s^2 \Theta(s, x, x)| < 6 \times \log(7/5)^2 \times (7/5)^{-s} < (7/5)^{-s}. \quad (19.19)$$

Define

$$s_k = 16 + k/4. \quad (19.20)$$

When $k = 32$ we have $s_k = 24$. We compute explicitly in Mathematica that

$$-(7/5)^s \times \Theta\left(s, 1 - \frac{2}{s}, 1 - \frac{2}{s}\right) - \frac{1}{64} > 0, \quad (19.21)$$

$$-(7/5)^s \times \Theta\left(s, 1 - \frac{2}{s - (1/4)}, 1 - \frac{2}{s - (1/4)}\right) - \frac{1}{64} > 0, \quad (19.22)$$

for the 33 values $s = s_0, \dots, s_{32}$. In both cases, the min is greater than .005 and so this calculation only requires 3 digits of accuracy.

Consider the interval $[s_k, s_{k+1}]$ for $k \in \{0, \dots, 31\}$. The calculation above shows that

$$\begin{aligned} & \Theta\left(s_k, 1 - \frac{2}{s_k}, 1 - \frac{2}{s_k}\right), \Theta\left(s_k, 1 - \frac{2}{s_{k+1}}, 1 - \frac{2}{s_{k+1}}\right) < \\ & -\frac{1}{64}(7/5)^{-s} < -\frac{1}{128}(7/5)^{-s} = -\frac{1}{8} \times (7/5)^{-s} \times \left(\frac{1}{4}\right)^2 < \\ & -\frac{1}{8} \max_{s \in [s_k, s_{k+1}]} \left| \partial_s^2 \Theta\left(s, 1 - \frac{2}{s}, 1 - \frac{2}{s}\right) \right| \times |s_{k+1} - s_k|^2. \end{aligned} \quad (19.23)$$

But then, by Lemma 16.2.1, applied to $-\Theta$,

$$\Theta\left(s, 1 - \frac{2}{s}, 1 - \frac{2}{s}\right) < 0, \quad \forall s \in [s_k, s_{k+1}]. \quad (19.24)$$

We get this result for all $k = 0, \dots, 31$, and so we conclude that

$$\Theta\left(s, 1 - \frac{2}{s}, 1 - \frac{2}{s}\right) < 0, \quad \forall s \in [16, 24]. \quad (19.25)$$

This proves that the TBP is not a minimizer w.r.t R_s for any $s \in [16, 24]$.

19.6 The End

Letting $x_s = 1 - 2/s$, we compute

$$A(s, x_s) = \left(\frac{2 - 2s + s^2}{4(s - 2)^2 s^2} \right)^{s/2} < \left(\frac{1}{4} - \frac{1}{20s} \right)^{s/2}. \quad (19.26)$$

Our estimate holds for all $s \geq 22.0453\dots$ We prove this by setting the two quantities inside the brackets equal to each other, solving for s , and noting that all real roots are less than $22.0453\dots$ This shows that the difference between the two quantities does not change sign after this value. A single evaluation at $s = 24$ confirms that the inequality goes in the correct direction.

Next, we compute that

$$B(s, x_s) = \left(\frac{1}{2} - \frac{1}{s} + \frac{1}{s^2} \right)^{s/2} \leq \left(\frac{1}{2} - \frac{23}{24s} \right)^{s/2}. \quad (19.27)$$

Using the same method, we establish that this estimate holds for all $s \geq 24$.

Finally, we compute

$$C(s, x_s, x_s) = \left(\frac{2 - 2s + s^2}{2(s - 2)^2 s^2} \right)^{s/2} \leq \left(\frac{1}{2} - \frac{1}{10s} \right)^{s/2} \quad (19.28)$$

This inequality follows from Equation 19.26 and holds in the same range.

We take $s \geq 24$ so that all the inequalities above hold. From Equation 16.7 we have

$$R(s, x_s, x_s) = 2^{-s/2}(a + b + c), \quad (19.29)$$

where

$$a = 2 \left(\frac{1}{2} + \frac{1}{10s} \right)^{s/2} < 0.01. \quad (19.30)$$

$$b = 4 \left(1 - \frac{23}{12s} \right)^{s/2} \leq 4 \exp(-23/24) < 1.54. \quad (19.31)$$

$$c = 4 \left(1 + \frac{1}{5s} \right)^{s/2} \leq 4 \exp(1/10) < 4.43. \quad (19.32)$$

Equation 19.30 is true by a mile. Equations 19.31 and 19.32 follow from the well known fact that

$$\lim_{t \rightarrow \infty} \mu(t) = \exp(k), \quad \mu(t) = \left(1 + \frac{k}{t} \right)^t, \quad (19.33)$$

and that $\mu(t)$ is monotone increasing on $(|k|, \infty)$.

Adding up these estimates, we get

$$R_s(s, x_s, x_s) < 5.98 \times 2^{-s/2} < 6 \times 2^{-s/2} + 3 \times 3^{-s/2} + 4^{-s/2} = R_s(TBP). \quad (19.34)$$

This proves that the TBP is not a minimizer w.r.t. R_s for $s \in [24, \infty)$.

Our proof of the Main Theorem is done.

References

[A] A. N. Andreev, *An extremal property of the icosahedron* East J Approx **2** (1996) no. 4 pp. 459-462

[BBCGKS] Brandon Ballinger, Grigoriy Blekherman, Henry Cohn, Noah Giansiracusa, Elizabeth Kelly, Achill Schurmann,
Experimental Study of Energy-Minimizing Point Configurations on Spheres,
arXiv: math/0611451v3, 7 Oct 2008

[BDHSS] P. G. Boyvalenkov, P. D. Dragnev, D. P. Hardin, E. B. Saff, M. M. Stoyanova, *Universal Lower Bounds and Potential Energy of Spherical Codes*, Constructive Approximation 2016 (to appear)

[BHS], S. V. Bondarenko, D. P. Hardin, E.B. Saff, *Mesh Ratios for Best Packings and Limits of Minimal Energy Configurations*,

[C] Harvey Cohn, *Stability Configurations of Electrons on a Sphere*, Mathematical Tables and Other Aids to Computation, Vol 10, No 55, July 1956, pp 117-120.

[CK] Henry Cohn and Abhinav Kumar, *Universally Optimal Distributions of Points on Spheres*, J.A.M.S. **20** (2007) 99-147

[CCD] online website:
<http://www-wales.ch.cam.ac.uk/~wales/CCD/Thomson/table.html>

[DLT] P. D. Dragnev, D. A. Legg, and D. W. Townsend, *Discrete Logarithmic Energy on the Sphere*, Pacific Journal of Mathematics, Volume 207, Number 2 (2002) pp 345–357

[Fö], Föppl *Stabile Anordnungen von Electron in Atom*, J. für die Reine Agnew Math. **141**, 1912, pp 251-301.

[HS], Xiaorong Hou and Junwei Shao, *Spherical Distribution of 5 Points with Maximal Distance Sum*, arXiv:0906.0937v1 [cs.DM] 4 Jun 2009

[I] IEEE Standard for Binary Floating-Point Arithmetic (IEEE Std 754-1985) Institute of Electrical and Electronics Engineers, July 26, 1985

[KY], A. V. Kolushov and V. A. Yudin, *Extremal Dispositions of Points on the Sphere*, Anal. Math **23** (1997) 143-146

[MKS], T. W. Melnyk, O. Knop, W.R. Smith, *Extremal arrangements of point and unit charges on the sphere: equilibrium configurations revisited*, Canadian Journal of Chemistry 55.10 (1977) pp 1745-1761

[RSZ] E. A. Rakhmanoff, E. B. Saff, and Y. M. Zhou, *Electrons on the Sphere*, Computational Methods and Function Theory, R. M. Ali, St. Ruscheweyh, and E. B. Saff, Eds. (1995) pp 111-127

[S1] R. E. Schwartz, *The 5 Electron Case of Thomson's Problem*, Journal of Experimental Math, 2013.

[S2] R. E. Schwartz, *The Projective Heat Map*, A.M.S. Research Monograph, 2017.

[S3] R. E. Schwartz, *Lengthening a Tetrahedron*, Geometriae Dedicata, 2014.

[SK] E. B. Saff and A. B. J. Kuijlaars, *Distributing many points on a Sphere*, Math. Intelligencer, Volume 19, Number 1, December 1997 pp 5-11

[Th] J. J. Thomson, *On the Structure of the Atom: an Investigation of the Stability of the Periods of Oscillation of a number of Corpuscles arranged at equal intervals around the Circumference of a Circle with Application of the results to the Theory of Atomic Structure*. Philosophical magazine, Series 6, Volume 7, Number 39, pp 237-265, March 1904.

[T] A. Tumanov, *Minimal Bi-Quadratic energy of 5 particles on 2-sphere*, Indiana Univ. Math Journal, **62** (2013) pp 1717-1731.

[W] S. Wolfram, *The Mathematica Book*, 4th ed. Wolfram Media/Cambridge University Press, Champaign/Cambridge (1999)

[Y], V. A. Yudin, *Minimum potential energy of a point system of charges* (Russian) Diskret. Mat. **4** (1992), 115-121, translation in Discrete Math Appl. **3** (1993) 75-81